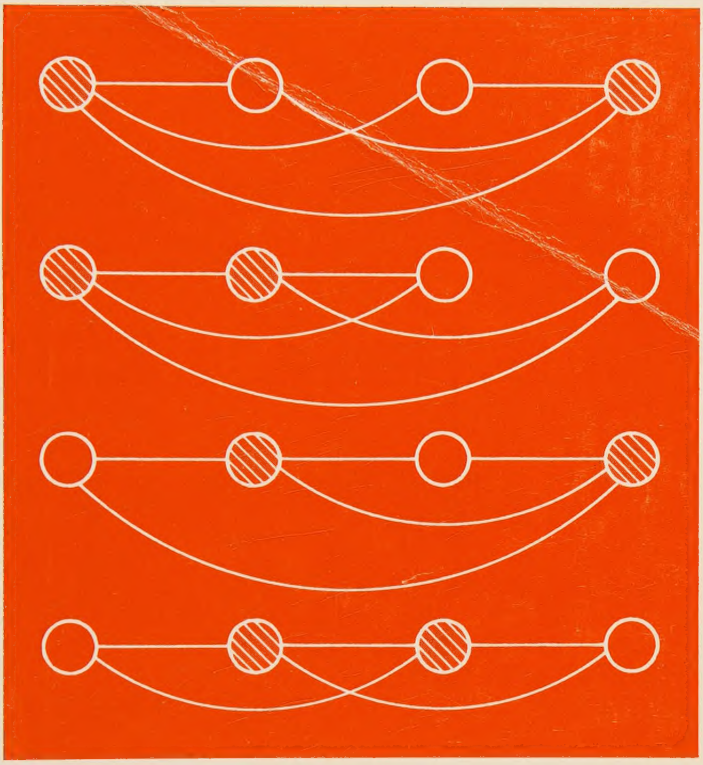


Statistical inference in graphs

1971

OVE FRANK

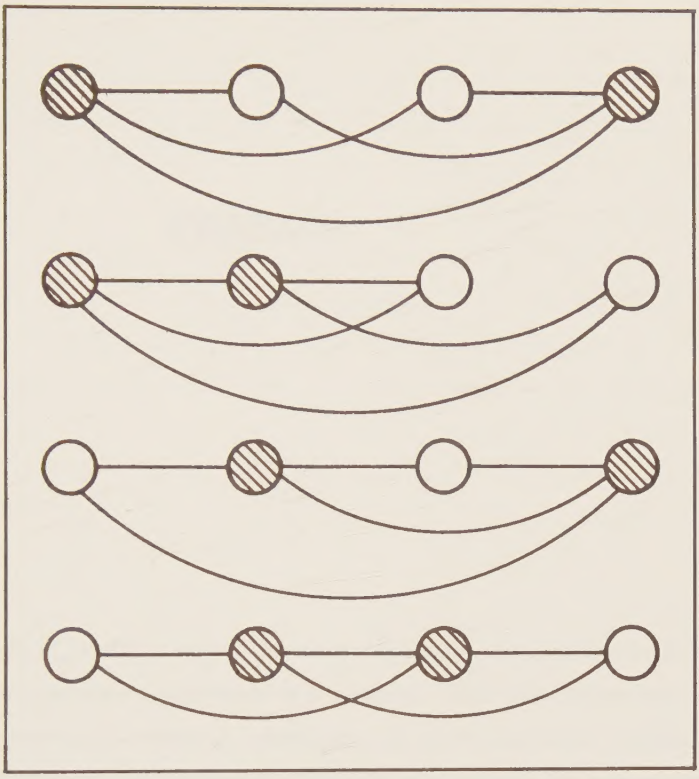
Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41552724_0023

Statistical inference in graphs

OVE FRANK

UNIV.
LEIDEN
BIBL.



Statistical inference in graphs

OVE FRANK

Printed by FOA Repro
Försvarets Forskningsanstalt
Stockholm 1971

PREFACE

"Systems analysis" is often used to denote methods for studying interrelations or connections between components in a complex. This wide-sense concept of a system comprises technical, biological, economic and social applications. In many applications the mathematical theory of graphs is the natural tool for describing and analysing the systems. In 1968 I started a research project with the aim to focus the attention on various statistical problems arising in connection with graph methods, and to initiate the development of a statistical systems theory.

The Tercentenary Fund of the Bank of Sweden provided a grant in support of my project "Systems Analysis with Probability and Graph Methods", in the course of which the work presented in this book was carried out. Some of the new results which are given in the book were originally obtained in my consulting work for the Research Institute of National Defence in Stockholm.

For kind support of the project I am greatly indebted to Professor Sten Malmquist, Professor Tore Dalenius, Dr. Carl-Gustaf Jennergren, and Professor Lars Erik Zachrisson. Professor Malmquist and Professor Dalenius have also contributed by directing my attention to various interesting fields of application and to many valuable references. Since many years, Dr. Gustaf Borenius and I have a common interest in the beautiful and compact tools provided by generali-

zed matrix inverses, and we had many interesting discussions about applications of these inverses to the flow-estimation problems which are treated in a chapter of this book. His great interest in my problems has been very stimulating.

The numerous figures occurring in the book were drawn by my wife Kerstin. Mr. Olov Alvfeldt revised the manuscript from a formal point of view and organized the printing, Mr. L.J. Gruber checked the language, Mrs. Git Sundt typed the final text, and Mrs. Anna-Lisa Persevall typed the manuscript. I am very grateful to them all.

CONTENTS

CHAPTER 1. INTRODUCTION 13

Synopsis 13

1. Applied mathematics 14
1.1 The use of mathematical methods 14
1.2 Construction of models 15
1.3 Development of methodology 16

2. Graph methods 17
2.1 The concept of a graph 17
2.2 The evolution of graph theory 18
2.3 A comment on terminology 20

3. The purpose and disposition of the book 20
3.1 General purpose 20
3.2 Disposition and special purposes 21

CHAPTER 2. GRAPHS 23

Synopsis 23

1. Basic concepts 25
1.1 Introduction 25
1.2 Directed graphs 26
1.3 Undirected graphs 27
1.4 The structure concept 28

2. Connectedness 29
2.1 Adjacency 29
2.2 Paths 30
2.3 Connected components 34
2.4 A connectedness hierarchy 35

3. Distance 36
3.1 Definition and simple properties 36
3.2 Determination of the distance matrix 37
3.3 Centrality 38
3.4 Generalization to valued graphs 40

4. Capacity 41
4.1 Introduction 41
4.2 Path frequencies 41
4.3 Elementary path frequencies 42
4.4 The maximum number of disjoint paths 44

5. Independence and dominance 47
5.1 Independence 47
5.2 Chromatic decompositions 48
5.3 Dominance 49

CHAPTER 3. STOCHASTIC GRAPHS 51

Synopsis 51

1. Introduction 53
1.1 The concept of a stochastic graph 53
1.2 Combinatorial graph problems 53
1.3 Three basic generators of stochastic graphs 54

2. Sampling models 55
2.1 Nexus sampling 55
2.2 Snowball sampling 56
2.3 Sampling social networks 60
2.4 Sampling and observation procedures in a graph 60

3. Measurement error models 63
3.1 Scheduling problems 63
3.2 Flow problems 64

4. Randomization models 65
4.1 Random arcs 65
4.2 Random intersection of two graphs 66
4.3 Random directions 67

CHAPTER 4. INFERENCE FROM SAMPLED SUBGRAPHS 68

Synopsis 68

1. The population graph 70
1.1 Preliminaries 70
1.2 The population graph 71

2. The sample graph 72
2.1 The sampling distribution of subgraphs 72
2.2 Notations 73
2.3 A graph list 75
2.4 Characteristics 76
2.5 Example 1 76

2.6 Example 2 80
2.7 The sample selection indicators 86
2.8 Stochastic properties of the arc frequencies 86

3. Estimation of the arc frequency 89
3.1 Introduction 89
3.2 An unbiased estimator and its variance 91
3.3 An unbiased estimator of the variance 92
3.4 A numerical example 94
3.5 Application to a problem in sociometry 95

4. Estimation of the node and arc totals 96
4.1 Introduction 96
4.2 A general sampling scheme 97
4.3 Unbiased estimators and their variances 99
4.4 Parametrization of the population graph 101
4.5 Unbiased estimators of the variances 103
4.6 Estimation of the node total 104
4.7 Estimation of the arc total 105
4.8 Estimation of the sum of the node and arc totals 106
4.9 A priori knowledge about the population graph 107
4.10 A matrix sampling design 107
4.11 A comparison between matrix sampling and ordinary sampling 108

5. Estimation of the distribution of the local degrees 109
5.1 Introduction 109
5.2 The undirected case 109
5.3 A comment about the covariances 112
5.4 The directed case 113
5.5 Application to a problem in sociometry 115

6. Estimation of the number of connected components 117
6.1 Notations 117
6.2 Complete components 119
6.3 Sampling without replacement 121
6.4 Bernoulli sampling 123
6.5 An example 126
6.6 Tree components 127
6.7 Comments 129

CHAPTER 5. INFERENCE FROM SAMPLED PARTIAL GRAPHS 130

Synopsis 130

1. Sampled partial graphs 132
1.1 A partial graph obtained by node sampling 132
1.2 Generalization to directed graphs 133
1.3 An example of a sampling distribution of partial graphs 133

2. Estimation of the arc frequency 134
2.1 The undirected case 134
2.2 Observation of out-arcs 137
2.3 Observation of out-arcs and in-arcs 138
2.4 Comparisons between subgraph and partial graph inference 138

3. Estimation of the node and arc totals 140
3.1 Introduction 140
3.2 Unbiased estimators and their variances 140
3.3 Unbiased estimators of the variances 142

4. Estimation of the distribution of the local degrees 143
4.1 The undirected case 143
4.2 The directed case 144
4.3 A numerical example 146
4.4 Comments about applications 149

5. Estimation of the number of connected components 150
5.1 Introduction 150
5.2 Complete components 151

CHAPTER 6. INFERENCE FROM FLOW MEASUREMENTS 152

Synopsis 152

1. Introduction 154
1.1 Flows in graphs 154
1.2 Log floating 155
1.3 Mail volumes 156

2. A model of flow measurement 156
2.1 The general problem 156
2.2 Notations 157
2.3 Least squares estimators 158
2.4 Repeated measurements 159
2.5 General assumptions 160
2.6 Comments 161

3. Application of linear theory 162
3.1 Preliminaries 162
3.2 Notations 165
3.3 Admissible node flows 166
3.4 Consistent flows 167
3.5 Unique arc flows 167
3.6 An example 168
3.7 Least squares estimators 171
3.8 Properties of the estimators 173

4. Flow estimation in complete graphs 175
4.1 Complete graphs 175

4.2 Explicit solutions 176
4.3 Application to a problem of class transitions 178

5. Flow estimation in complete bipartite graphs 179
5.1 Bipartite graphs 179
5.2 Complete bipartite graphs 180
5.3 Explicit solutions 181
5.4 Application to two-way contingency tables 183

6. Flow estimation in trees 186
6.1 Trees 186
6.2 The terminal node classes 187
6.3 The method of solution 189
6.4 The induction step 190
6.5 The central solution 191
6.6 The algorithm 192
6.7 An example 192
6.8 Generalization 194
6.9 Notes 195
6.10 Regular trees 196
6.11 A non-optimal estimator 198
6.12 Application to nested classifications 199

CHAPTER 7. INFERENCE FROM STOCHASTIC CONTACT GRAPHS 201

Synopsis 201

1. Stochastic contact graphs 203
1.1 Arc sampling 203
1.2 Random choice of contacts 204

2. Undirected complete population graphs 205
2.1 Introduction 205
2.2 Basic properties 205
2.3 Fixed number of contacts 207
2.4 The isolate problem of sociometry 207
2.5 The clique problem of sociometry 209
2.6 Two problems of reliability 211
2.7 Modifications when circuit arcs are present 212

3. Directed complete population graphs 213
3.1 Introduction 213
3.2 Basic properties 214
3.3 Fixed number of contacts 216
3.4 Fixed local out-degrees 217
3.5 The reciprocal choice problem of sociometry 218
3.6 The isolate problem of sociometry 221
3.7 The clique problem of sociometry 224
3.8 Modifications when circuit arcs are present 226

4. General population graphs 226
4.1 The undirected case 226
4.2 Estimation of the distribution of the local degrees 227
4.3 The isolate problem 228
4.4 An application to quality control 229
4.5 The directed case 230

CHAPTER 8. INFERENCE FROM STOCHASTIC PREFERENCE GRAPHS 233

Synopsis 233

1. Ranking based on paired comparisons 235
1.1 Paired comparisons and preferences 235
1.2 Comparison graphs and preference graphs 236
1.3 Ranking procedures 237

2. Stochastic preference graphs 239
2.1 Preference probabilities 239
2.2 A preference model 240
2.3 Some illustrations of the preference model 241

3. Complete preference graphs 242
3.1 Notations 242
3.2 The scores 243
3.3 Agreement 245
3.4 Inconsistency 248
3.5 Relationship between agreement and inconsistency 252
3.6 Complete orders 253
3.7 Estimation of the preference probability 255

4. Complete bipartite preference graphs 258
4.1 Introduction 258
4.2 Agreement 259
4.3 Inconsistency 259

REFERENCES 263

AUTHOR INDEX 273

SUBJECT INDEX 277

Chapter 1

INTRODUCTION

SYNOPSIS

1. Applied mathematics. The increasing use of applied mathematics is stressed. The development of automatic computer technology, mathematical methodology and model-thinking are considered. The construction of models and the development of mathematical methods are discussed, and the alternative contributions made to this evolution by new fields of application and new theory are particularly emphasized.

2. Graph methods. The concept of a graph is introduced, and some basic kinds of graph are defined. The connection with binary relations is considered. An account of the evolution of graph theory is given together with references to the literature.

3. The purpose and disposition of the book. The book treats statistical methods in graph models and aims at stimulating further research about, and increased application of, such methods. Three main types of models with stochastic graphs are used in the book. They may be called sampling models, measurement error models and randomization models.

1. APPLIED MATHEMATICS

1.1 The use of mathematical methods

Mathematical methods have gained increased importance in more and more fields of application. By the development of computer science and information processing systems it has become possible to tackle problems involving larger and larger amounts of computation. Algorithms have been developed to solve complicated problems that are not accessible to an approach based upon the methods of classical mathematics. Besides the rapid growth of computer techniques and computer capacity an extensive evolution of methodology can be noted. The consciousness that mathematical methods are useful not only for readily identified computation problems but also for description, analysis and optimization of large systems of a technical, economical, social, biological or other nature has resulted in a great expansion of applied mathematics in various sciences. Econometry, sociometry, psychometry, biometry and so forth may be mentioned as examples of established terms of methodology in various fields of application.

The concept of a mathematical model has come to play an important role in applied mathematics. When a real phenomenon is described in mathematical terms certain facts and properties are emphasized and certain others are suppressed. One obtains a model of reality. Owing to the evolution of methodology which has taken and is taking place, a great variety of different tools have become available to construct and treat mathematical models.

Before we turn to a study of the use of graph methods in model construction we will emphasize two general aspects of the concept of a model. The first aspect concerns the requirement for both methodological knowledge and a good understanding of the practical problem in order that construction of a good model should be possible. The second aspect concerns the interaction between theory and application as a basis for the development of the mathematical methods. The next two subsections will be devoted to these two aspects.

1.2 Construction of models

A prerequisite of successful model construction is a thorough knowledge of the various methods and solution techniques. The appropriate formulation of the problems is then facilitated, and the choice of a model so complicated that it can hardly by analyzed may be avoided. The problems can be posed in such a way that the mathematical treatment becomes tractable, and the model may be expected to give results of practical value with a reasonable amount of work.

It is obvious that a thorough knowledge of methodology makes it possible to economize on mental work and to identify easily any standard method appropriate for a practical problem. It is also obvious, however, that an uncritical use of standard methods involves certain risks. The problems may be corrupted to fit a desired method. A comprehensive and unrestricted analysis of the problems may be obstructed by a choice of routine methods only.

A practical problem may be given in mathematical terms in several alternative ways, and different mathematical models may be more or less successful when used to describe and analyze the problem. The choice of a model is determined by the way one looks at the practical problem. If this way is not appropriate or if it is wrong in any aspect, this may mean that the studied mathematical problem is peripheral or irrelevant to the actual application. However advanced the methods may be, they are not capable of compensating for a wrongly posed problem. It may be said that a good model is essentially based upon good knowledge about the reality one wants to describe and analyze by the model. Without such knowledge one is exposed to the risk of missing important components in the problem or of placing too much emphasis on other less important components.

If a good knowledge of the problem but a limited mathematical view are used to tackle the problem, the choice of model may become unnecessarily poor, resulting in a much too superficial treatment of the problem. The conclusion may then easily be that mathematical methods are of no use.

As a consequence it is of fundamental importance that the choice of a mathematical model is based upon thorough knowledge of both the practical problem

and the available and appropriate methods. These two kinds of required knowledge are often possessed by different persons, and then their ability to communicate and cooperate is important in order to achieve good treatment of the problem.

1.3 Development of methodology

The arsenal of methods that is available to construct and analyze mathematical models is augmented and changed by the influence of theory and application.

The effects of application on the development of methodology result from practical problems that are posed in mathematical terms. If the models obtained cannot be analyzed by known methods, or if a certain method cannot be applied directly, there is need of new methods or a modification of an old one. In this way various fields of application may recommend how the methodology ought to be developed to fit particular problems.

The theoretical development of methodology starts from a certain methodological problem or its theoretical basis. By formalization and simplification, related problems are obtained which may be interpreted in alternative ways and tackled by different methods. Comparisons and analogies with other problems may be carried out, and several possible applications may be considered. By formulating the problems abstractly, common elements and further possibilities to generalize and formalize may be found. In this way theoretical research problems are created whose solutions may contribute to the development of methodology in several fields of application.

The theoretical development of methodology suggests new fields of application. When the problems of these fields of application are studied they may provide the impetus for further development of methodology. This alternating effect of theory and application is fundamental to all applied research.

2. GRAPH METHODS

2.1 The concept of a graph

One of the several mathematical tools that are available for model building is graph methodology. Graphs have been systematically studied by König already in the thirties, but the essential evolution of graph models belongs to the sixties.

A graph may be described simply as a set of objects together with a relationship that is satisfied by certain pairs of objects. In a figure the objects can be represented by nodes and the related objects connected by arcs. To make the graph concept concrete a few examples may be given.

The results of team competitions can be illustrated by a graph where the nodes represent the teams and the arcs represent the relation "defeated". A genealogical table can be given as a graph where the nodes represent the individuals and the arcs mean "parent to". An economical or technical system may be described by a graph where the nodes correspond to the states of the system and the arcs represent the possible transitions between the states.

The arcs of a graph can be directed or undirected. There may be arcs starting and ending at the same node, i.e. circuit arcs. There may be several arcs between the same pair of nodes, i.e. multiple arcs. The nodes and the arcs may be associated with various characteristics or measurements, i.e. there may be qualitative or quantitative variables defined in the node set and the arc set.

Graphs suitable for various fields of application may be obtained by specifying the properties of the graphs in different ways. Some basic types of graph are defined by use of the concepts mentioned above. Simple graphs have no circuit arcs, no multiple arcs and no node or arc values. Directed graphs have all their arcs directed, and undirected graphs have all their arcs undirected. Valued graphs are associated with node or arc values.

Mathematically a graph can be said to consist of two abstract sets, called the node set and the arc set, and a function which associates a unique pair of nodes with each arc. Ordered pairs of nodes correspond to directed graphs, and

unordered pairs of nodes correspond to undirected graphs. If the function is unique both ways, so that each pair of nodes is associated with at most one arc, i.e. if multiple arcs are not present, the set of arcs can be considered to be a subset of the Cartesian product of the node set with itself, and the graph can be said to represent a binary relation.

The terminology of binary relations is easily carried over to the more visual graph terminology. A reflexive relation corresponds to a graph with a circuit arc at every node. An irreflexive relation corresponds to a graph without circuit arcs. A symmetrical relation corresponds to a graph with an oppositely directed arc to every arc, or equivalently an undirected graph. A transitive relation corresponds to a graph in which each pair of nodes that are connected by a chain of equally directed arcs are also connected by an arc in the same direction.

2.2 The evolution of graph theory

Many early applications of graph theory belong to the category of puzzle problems and mathematical recreations. The famous Königsberg Bridge Problem considered by Euler in 1736 belongs to this category. So does also probably the most celebrated one of all such popular problems, the Four Color Map Conjecture, posed by De Morgan in 1852, which is as yet unsolved and the object of many people's interest. The numerous contributions to graph theory that have been developed for this problem are presented in a book by Ore (1967).

In the middle of the nineteenth century some methods of graph character began to be used in the natural sciences. Kirchhoff studied electrical circuits by the use of methods that are precursors to the modern network theory. In chemistry, Cayley considered structure problems for crystals and molecules by the use of methods that may nowadays be called combinatorial graph theory.

In the beginning of the twentieth century several mathematicians were using graphs as tools, for instance Menger in topology and Hertz in formal logic. The name "graph" and the systematical treatment of various properties of graphs from a pure mathematical point of view were introduced by König (1936).

Due to the progress of operations research and the development of game theory and linear programming, graph theory has gained current interest. Problems of transportation, allocation, assignment and so forth are all typical operations research problems, and such problems in particular have been solved by the use of algorithms based on linear programming, integer programming or graph theory. Some references are Gale (1960), Ford and Fulkerson (1962), Avondo-Bodino (1962), Danzig (1963), and Berge and Ghouila-Houri (1965).

About 1960 two related management techniques known as CPM (Critical Path Method) and PERT (Program Evaluation and Review Technique) were developed and applied in operations analysis. Two early references are Malcolm et al. (1959) and Kelley (1961). CPM and PERT are designed for planning and scheduling of large projects composed of many interacting components. By the use of graph theory several algorithms have been invented to solve various types of cost and time scheduling problems. References may be made to Fulkerson (1961), Grossman and Lerchs (1961), and Berman (1964), to mention only a few papers emphasizing graph methods.

In electrical science many of the methods used to analyze switching circuits are based on graph theory. The electrical applications have contributed much to the evolution of graph theory, and they still do influence it considerably. Some books about electrical network theory are those by Cochrun (1967), Krieger (1967), and Blackwell (1968).

During the sixties an increased use of mathematical models can be noted in the social and behavioral sciences. The methods of studying relations between objects that are proposed by graph theory have found frequent applications in the studies of social behavior between human beings, families, groups, etc. The problems of social science have contributed to the development of particular parts of graph theory. The social science graph problems have been treated by Flament (1963) and Harary, Norman and Cartwright (1965).

Graphs have been used as descriptive tools in information theory, automata theory and linguistics, and many problems in these subjects have been given as complicated combinatorial graph problems. Gill (1962), Ginsburg (1962), Floyd

(1963), Salomaa (1969), and Chomsky (1968) may be referred to.

A large proportion of the graph theory literature is associated with particular fields of application. General graph theory, however, is also treated in some mathematical works of a more up-to-date appearance than the original work by König. Berge (1958), Ore (1962), Hammer and Rudeanu (1968), and Roy (1969) give mathematical accounts of graph theory.

The mathematical content of graph theory is closely related to Boolean algebra. Many results are of a combinatorial nature, and many graph procedures are of an algorithmical type. The rapidly increasing interest in algorithmical solution methods which is a consequence of the evolution of automatic computer techniques has brought about an expansion of discrete finite mathematics. This is probably a field that will influence very greatly the future development of graph theory. In this connection it may be mentioned that an interesting outline of an analysis of algorithms has been presented by Knuth (1968) in the introductory volume of an extensive seven-volume set of books.

2.3 A comment on terminology

The strong connection between graph theory and many different fields of application has resulted in terminology becoming somewhat inconvenient and entangled. The various fields of application have created their own notations and definitions. No generally accepted terminology exists even in mathematical books about graph theory. With the exception of a few basic concepts, however, the need of a unified terminology is rather slight. Too many definitions appear in a great deal of the literature. Many concepts are defined which are not needed to derive useful results but are rather used to analyze the connections between a lot of redundant concepts.

3. THE PURPOSE AND DISPOSITION OF THE BOOK

3.1 General purpose

The general purpose of this book is to introduce statistical graph problems and to indicate some potential applications in order to stimulate research about,

and the use of, stochastic graphs.

The discussion of the development of methodology in Subsection 1.3 emphasized the interaction between theory and application. The knowledge about methods may provide the stimulus for applications in various fields, and inducements to further development of methodology may be obtained from the applications. This is the background which should be remembered when some of the statistical graph problems encountered in the following chapters are incompletely solved and badly in need of further working to be of value to the applications.

The problems that will be considered are all inspired by concrete applications which the author has come across in his consulting work. In order that the problems should be capable of treatment, however, they have not always been kept in their original version but have sometimes been simplified by particular assumptions. Generally speaking, the statistical graph problems obtained have not been treated before, and it has therefore been considered appropriate to present also such results that may inspire further research, even if they are at present of limited applicability.

3.2 Disposition and special purposes

An outline of the contents of the book will be given here. More detailed accounts of the contents of each chapter are found in the synopsis at the beginning of each chapter.

Chapter 2 is a survey of fundamental concepts of graph theory. The account aims at making the reader accustomed to the methods of analysis that can be used for graphs. Some graph theory results which are presumed to be of general interest for further research about statistical graph problems are presented.

Chapter 3 contains a systematic classification of stochastic graphs intended to be useful for the identification of problems that can be tackled by various kinds of graph methods. A division is made into three types of stochastic graph, and particular representatives of these three types of graph are studied in Chapter 4 to 8.

The first type of stochastic graph is obtained by application of some random

sampling and observation procedures which provide information about a part of a graph, i.e. a sample graph. Stochastic graphs of this type are the basis of a theory of sampling and statistical inference pertinent to populations with a graph structure, i.e. populations whose units are related in pairs. To contribute to the development of such a theory two different observation procedures are studied in Chapter 4 and 5. Chapter 4 deals with various inference problems which occur when the observed sample graph is a subgraph of the population graph. Chapter 5 deals with the analogous inference problems that occur if the observed sample graph is a partial graph of the population graph. There are many potential applications of a theory of graph inference for the studied sampling and observation procedures.

The second type of stochastic graph is obtained by the introduction of uncertainty into the node or arc variables that are defined in a valued graph. This type of stochastic graph is the basis of measurement error models with graphs. In Chapter 6 flow estimation problems in graphs are studied to illustrate the measurement error models. The arc flows are assumed to be observed with measurement errors. The inference from the observed to the real arc flows is considered. Several examples are given to show the applicabilities of flow measurement and inference.

The third type of stochastic graph is obtained by randomized graph construction and can be illustrated by graphs with stochastically appearing arcs or stochastically directed arcs. Such graphs are applicable in sociometry, in reliability theory, and in the theory of paired comparisons. Some simple stochastic models with such graphs are considered in Chapter 7 and 8.

Chapter 2

GRAPHS

SYNOPSIS

1. Basic concepts. The basic concepts of graph theory are introduced in a way that is slightly more general and formally simpler than that met with in the standard texts. This is achieved by a consequential use of the symbolism of set theory.

2. Connectedness. In order to study the connectedness of the nodes of a graph, adjacency operators and arc indicators are introduced. Paths are defined, and their occurrence is determined by an algorithmical procedure. A method is given to decompose a graph into its connected components and its possibly occurring arcs between the components. Finally, a hierarchical division of the node set into three kinds of connected classes is described.

3. Distance. The distance matrix is defined, and some simple properties of the matrix are described. An algorithm is given to compute the distance matrix. The centrality properties of graphs are discussed. A generalization to valued graphs is indicated.

4. Capacity. An introductory discussion is devoted to how the number of paths with certain specified properties may be interpreted as measures of ca-

pacity in various applications. Algorithms are provided that give the number of paths and elementary paths and the maximum number of disjoint paths between the different pairs of nodes.

5. Independence and dominance. The way in which various applications may raise the interest in node subsets with certain specified properties is indicated by the use of examples. Independent node sets, and particularly their occurrence in chromatic decompositions, are discussed. Dominating and dominated node sets are also dealt with.

1. BASIC CONCEPTS

1.1 Introduction

It will be clear from the discussion in Chapter 1, Subsection 2.1, that the concept of a graph can be defined in very general terms and made to include both directed and undirected arcs. There may be circuit arcs and multiple arcs. Moreover, the nodes and the arcs may be associated with certain properties or numerical values, thereby providing what is called a valued graph.

When a graph is analyzed, different aspects of the properties of the graph are emphasized depending on the real situation which is to be studied by the graph. It is therefore possible to specify the kind of graph in different ways so as to obtain one which is convenient for the analysis that is desired. Two simple examples will be given to concretize the choice of a suitable graph.

The arcs may for instance be of different strength or capacity, but in the problem at hand this can be of secondary importance. It may then be preferable to consider the graph without its arc values.

If the identities of the arcs are unimportant, and the only thing of interest is to know between which nodes the arcs occur, then it is possible to regard the multiple arcs as a single arc associated with a multiplicity number that denotes the number of multiple arcs which the single arc represents. In this way the graph with multiple arcs is replaced by a valued graph without multiple arcs.

Some methods which are convenient for analyzing various properties of graphs will be described in this chapter. The methods of analysis that are considered are all of a basic kind and occur frequently in the applications. Some types of problem that are dealt with can be described briefly as problems concerning the occurrence, the length and the number of different kinds of connections between the nodes, and problems which involve division of the node set into subsets of certain specified properties.

Before we pass on to the methods of analysis some basic definitions and notations of graph theory will be introduced. It may be noted that the account given here is formally somewhat different from the texts usually met with. By making a systematical use of both node and arc sets a general setup is achieved

which includes multiple arcs and allows many concepts to be defined by simple use of set theory.

1.2 Directed graphs

A directed graph consists of a node set Ω and a family of disjoint arc sets M_{ij} for $i \in \Omega$ and $j \in \Omega$. If the node set is made up of N nodes it is convenient to label the nodes by the natural numbers, i.e. to take $\Omega = \{1, 2, \ldots, N\}$. Each arc that belongs to M_{ij} is said to go from node i to node j and to have the start node i and the end node j. The set M_{ij} may consist of not one, exactly one, or more than one arc. If M_{ij} consists of only one arc this arc is called a single arc from node i to node j. If M_{ij} consists of several arcs they are called multiple arcs from node i to node j. The arcs belonging to M_{ii} are called circuit arcs at node i. The arcs belonging to M_{ij} are called out-arcs from node i and in-arcs to node j if $i \neq j$. The total arc set of the graph is

$$M = \bigcup_{i \in \Omega} \bigcup_{j \in \Omega} M_{ij} . \tag{1}$$

Figure 1 shows a diagram representing a graph with the node set $\Omega = \{1, 2, 3\}$ and the arc sets $M_{11} = \{1\}$, $M_{12} = \{2\}$, $M_{13} = \emptyset$, $M_{21} = \{3, 4\}$, $M_{22} = \emptyset$, $M_{23} = \emptyset$, $M_{31} = \emptyset$, $M_{32} = \{5, 6\}$, and $M_{33} = \{7\}$.

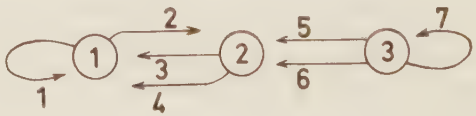


Figure 1

A graph with the node set $\Omega = \{1, 2, \ldots, N\}$ has an $N \times N$ arc indicator matrix A whose elements A_{ij} are equal to 1 or 0 according to whether node i has or has not at least one arc to node j. The $N \times N$ arc frequency matrix C has elements C_{ij} giving the number of arcs from node i to node j. Particularly, the diagonal element C_{ii} gives the number of circuit arcs at node i. The number of out-arcs from node i is given by the sum of the off-diagonal elements in row

i. It is called the out-degree of node i and is denoted by $a_i = \sum_{j \neq i} C_{ij}$. The number of in-arcs to node j is given by the sum of the off-diagonal elements in column j. It is called the in-degree of node j and is denoted by $b_j = \sum_{i \neq j} C_{ij}$. Let a, b and c denote the $N \times 1$ vectors that have as components the local out-degrees, the in-degrees and the circuit arc frequencies respectively. The total arc frequency R of the graph becomes

$$R = \sum_{i=1}^{N} \sum_{j=1}^{N} C_{ij} . \tag{2}$$

If the identities of the arcs are immaterial the graph can be represented by the node set Ω and the arc frequency matrix C. If no multiple arcs are present the graph can be represented by the node set Ω and the arc indicator matrix A. In this case the arcs may be denoted by ordered pairs of nodes, and the arc set M then becomes a subset of the Cartesian product Ω^2 of the node set with itself. It is then obviously true that $R \leq N^2$, which can be improved to $R \leq N(N-1)$ if circuit arcs do not occur.

A subgraph of a graph with node set Ω and arc sets M_{ij} for $i \in \Omega$ and $j \in \Omega$ is defined as a graph with node set $\Omega' \subset \Omega$ and arc sets M_{ij} for $i \in \Omega'$ and $j \in \Omega'$.

A partial graph of a graph with node set Ω and arc sets M_{ij} for $i \in \Omega$ and $j \in \Omega$ is defined as a graph with node set Ω and arc sets $M'_{ij} \subset M_{ij}$ for $i \in \Omega$ and $j \in \Omega$.

1.3 Undirected graphs

An undirected graph consists of a node set Ω and a family of arc sets M_{ij} which satisfy $M_{ij} = M_{ji}$ but are otherwise disjoint for $i \in \Omega$ and $j \in \Omega$. Each arc that belongs to M_{ij} is said to go between node i and node j. The concepts of single arc, multiple arc, circuit arc, total arc set, arc indicator matrix, arc frequency matrix and subgraph are defined as in the previous subsection. The definition of a partial graph given there needs the additional clause $M'_{ij} = M'_{ji}$ for $i \in \Omega$ and $j \in \Omega$.

The arc indicator matrix A and the arc frequency matrix C associated with

an undirected graph are symmetrical, i.e. $A = A'$ and $C = C'$, where the prime sign denotes the transposed matrix. The local out-degrees and in-degrees coincide and are called simply the local degrees. The total arc frequency R of an undirected graph with node set $\Omega = \{1, 2, \ldots, N\}$ satisfies

$$R = \sum_{i \leq j} \sum C_{ij} . \tag{3}$$

If the graph has no multiple arcs it holds that $R \leq (N+1)N/2$ which can be improved to $R \leq N(N-1)/2$ if circuit arcs do not occur.

1.4 The structure concept

Consider the two graphs of Figure 2, which both have N=4 nodes and R=2 arcs. The arcs labeled 1 and 2 have been changed in the two graphs. If the arcs are considered to be non-distinguishable, i.e. if the arc labels are omitted, the two graphs of Figure 2 become identical.

The two graphs of Figure 3 have both N=4 nodes and R=2 non-distinguishable arcs. If the nodes are considered to be non-distinguishable, i.e. if the node labels are omitted, the two graphs of Figure 3 become identical.

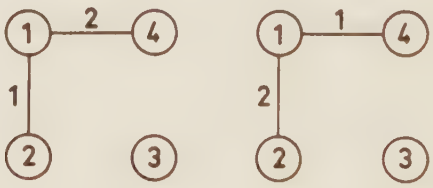


Figure 2

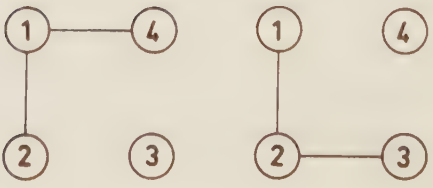


Figure 3

Graphs that become identical when their node and arc labels are omitted are said to have the same structure.

In order to determine whether two graphs represent the same structure their arc frequency matrices may be compared. Let C_1 and C_2 denote two arc frequency matrices. They represent the same structure if and only if there exists a permutation matrix P such that $P\,C_1\,P' = C_2$. By a permutation matrix is meant a matrix which consists of the digits 0 and 1, all row and column sums of which are equal to 1. The condition implies that the nodes of the first graph can be relabeled in such a way that the arc frequency matrices become identical.

2. CONNECTEDNESS

2.1 Adjacency

When studying the connectedness of the nodes in a graph it is possible to disregard the identities of the arcs and the multiple arcs. The basis of an analysis of connectedness is the adjacency properties of the nodes. The adjacency is conveniently described by use of two adjacency operators which will now be defined.

A directed graph with node set Ω and arc sets M_{ij} for $i \in \Omega$ and $j \in \Omega$ is considered. For each node $i \in \Omega$, a subset $G(i) \subset \Omega$ is defined which consists of all the nodes $j \in \Omega$ having arcs from node i. Thus

$$G(i) = \{j \in \Omega : M_{ij} \neq \emptyset\}. \tag{4}$$

Particularly it holds that $i \in G(i)$ if, and only if, there is a circuit arc at node i.

Moreover, for each node $j \in \Omega$, a subset $G'(j) \subset \Omega$ is defined which consists of all the nodes $i \in \Omega$ having arcs to node j. Thus

$$G'(j) = \{i \in \Omega : M_{ij} \neq \emptyset\}. \tag{5}$$

Obviously $j \in G(i)$ if, and only if, $i \in G'(j)$.

For undirected graphs the adjacency operators G and G' are formally defined as above. Since $M_{ij} = M_{ji}$, the position in this case is that $G(i) = G'(i)$ for $i \in \Omega$.

The adjacencies of a graph with the node set $\Omega = \{1, 2, \ldots, N\}$ are given by

the arc indicator matrix A according to

$$A_{ij} = \begin{cases} 1 \text{ if } j \in G(i), \\ 0 \text{ otherwise}, \end{cases} \tag{6}$$

for $i \in \Omega$ and $j \in \Omega$. The row sum $\sum_j A_{ij}$ denotes the number of nodes in G(i), i.e. the number of nodes which can be reached in one arc step from node i. The column sum $\sum_i A_{ij}$ denotes the number of nodes in G'(j), i.e. the number of nodes reaching node j in one arc step.

2.2 Paths

A path from node i to node j is a chain of equally directed arcs from node i to node j. More explicitly a path with s arc steps from node i_0 to node i_1 to node i_2 ... to node i_s is given by an arc sequence

$$(\alpha_1, \alpha_2, \ldots, \alpha_s), \text{ with } \alpha_v \in M_{i_{v-1} i_v}, \quad v = 1, 2, \ldots, s. \tag{7}$$

The path is said to contain the node sequence $(i_0, i_1, \ldots, i_s)$ and the arc sequence $(\alpha_1, \alpha_2, \ldots, \alpha_s)$. Particularly, if $i_0 = i_s$, the path is called a circuit path with s arc steps. A circuit path is said to contain the node sequence $(i_1, i_2, \ldots, i_s)$ and the arc sequence $(\alpha_1, \alpha_2, \ldots, \alpha_s)$. If no arc is contained in the path more than once, the path is called simple. If no node is contained in the path more than once, the path is called elementary. Consequently every elementary path is also simple. Every path from node i to node j contains a simple subpath from node i to node j, which in turn contains an elementary subpath from node i to node j.

If the identities of the arcs are disregarded, a path with s arc steps from node i_0 to node i_s can be given exclusively by a node sequence $(i_0, i_1, \ldots, i_s)$, with $i_v \in G(i_{v-1})$ for $v = 1, 2, \ldots, s$.

In Figure 4 the node sequence (1,2,5,4,2,3,5,2,3,6) represents a path from node 1 to node 6 which is not simple. A subpath from node 1 to node 6 which is simple but not elementary is given by the node sequence (1,2,5,4,2,3,6). An

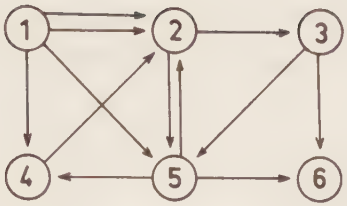


Figure 4

elementary subpath from node 1 to node 6 is given by the node sequence (1,2, 3,6).

In order to examine whether there is a path from node i to node j, use may be made of either the adjacency operator G or the arc indicator matrix A. These two methods will now be described.

Put

$$G(S) = \bigcup_{i \in S} G(i) \tag{8}$$

for arbitrary subsets $S \subset \Omega$. According to (8) we successively define

$$G_s(i) = G(G_{s-1}(i)) \tag{9}$$

for $s = 2,3,\ldots$ and $G_1(i)=G(i)$. The set $G_s(i)$ consists of the nodes j which can be reached in s arc steps from node i. Note that the sets $G_s(i)$ need not be disjoint for different values of s. The union set

$$\bigcup_{v=1}^{s} G_v(i) \tag{10}$$

consists of the nodes j which can be reached in at most s arc steps from node i. Particularly,

$$\hat{G}(i) = \bigcup_{v=1}^{N} G_v(i) \tag{11}$$

consists of the nodes which can be reached from node i in any number of arc steps. Thus $\hat{G}(i)$ provides information about the nodes which have paths from node i.

Let us now turn to the arc indicator matrix A. Define successively a matrix A_s with elements according to

$$(A_s)_{ij} = \max_{1 \le k \le N} A_{ik}(A_{s-1})_{kj} \tag{12}$$

for $s = 2,3,\ldots$ and $A_1 = A$. The matrix A_s is usually called the Boolean s-power of the matrix A. The matrix A_s is an indicator matrix of the paths with s arc steps, i.e. $(A_s)_{ij}$ is equal to 1 or 0 according to whether there is or is not at least one path with s arc steps from node i to node j. The matrix elements

$$\max_{1 \le v \le s} (A_v)_{ij} \tag{13}$$

are equal to 1 or 0 according to whether there is or is not at least one path with at most s arc steps from node i to node j. Particularly, the matrix elements

$$\hat{A}_{ij} = \max_{1 \le v \le N} (A_v)_{ij} \tag{14}$$

become equal to 1 or 0 according to whether there is or is not at least one path from node i to node j. The matrix $\hat{A}$ will be called the path indicator matrix.

By use of a slight trick, Maghout (1963) has been able to construct the path indicator matrix $\hat{A}$ in a way that is simpler than the one just described. To demonstrate the simpler method, introduce a matrix B with the elements

$$B_{ij} = \begin{cases} 1 & \text{if } i=j \text{ or if } j \in G(i), \\ 0 & \text{otherwise,} \end{cases} \tag{15}$$

i.e. B is obtained from the arc indicator matrix A by putting all the diagonal elements equal to 1. The matrix B_s defined in analogy with (12) has elements $(B_s)_{ij}$ which for $i=j$ are equal to 1 and for $i \neq j$ are equal to 1 or 0 according to

whether there is or is not at least one path with at most s arc steps from node i to node j. From this interpretation of the elements it follows that

$$B_{ij} \leq (B_2)_{ij} \leq \ldots \leq (B_{N-1})_{ij} = (B_N)_{ij} \tag{16}$$

for $i \in \Omega$ and $j \in \Omega$. Moreover, if there is a value $s = 1, 2, \ldots, N-1$ such that $B_s = B_{s+1}$, then it is true that $B_s = B_N$. The matrix B_N will be denoted by $\hat{B}$. Thus $\hat{B}$ is the first repeated matrix in the sequence $B_1, B_2, \ldots, B_N$, and

$$\hat{B}_{ij} = \begin{cases} 1 & \text{if } i=j, \\ \hat{A}_{ij} & \text{if } i \neq j, \end{cases} \tag{17}$$

i.e. $\hat{B}_{ij}$ indicates whether there is a path from node i to node j for $i \neq j$. A rapid way of determining $\hat{B}$ is to compute the sequence $B_1, B_2, B_4, B_8, \ldots$ of successive Boolean squares until the first repeated matrix is found. From $\hat{B}$ the circuit path indicators can be determined according to

$$\hat{A}_{ii} = \begin{cases} 1 \text{ if } A_{ii}=1, \\ \max\limits_{j \neq i} \hat{B}_{ij} \hat{B}_{ji} \text{ if } A_{ii}=0, \end{cases} \tag{18}$$

and then the path indicator matrix $\hat{A}$ is completely known.

Another simple algorithm to determine the path indicator matrix $\hat{A}$ has been described by Roy (1962). The algorithm is based on a series of successive transitivity completions at every node $k \in \Omega$, i.e. an arc from node i to node j is inserted if there is an arc from node i to node k and another arc from node k to node j. An operator T_k acting on arbitrary matrices is defined by

$$(T_k A)_{ij} = \max \{ A_{ij}, A_{ik} A_{kj} \} \tag{19}$$

for $i \in \Omega$ and $j \in \Omega$. The path indicator matrix is obtained as

$$\hat{A} = T_N T_{N-1} \cdots T_1 A. \tag{20}$$

It is practical to perform the transitivity completions successively in the same scheme. The initial scheme is A with an empty space substituted for the digit 0 and a cross mark substituted for the digit 1. The empty spaces are then filled up with the digit k, if a path is obtained when the transitivity completion is applied at node k. For a more detailed account of the method reference may be made to Hammer and Rudeanu (1968).

2.3 Connected components

For undirected graphs two different nodes are said to belong to the same connected class if there is a path between them. Thus the nodes i and j ($i \neq j$) are in the same connected class if $j \in \hat{G}(i)$, i.e. if $\hat{A}_{ij} = 1$. The nodes in a connected class and the arcs occurring between them constitute a connected component of the graph. The connected components of an undirected graph are subgraphs of the graph. An undirected graph is called connected if it consists of only one connected component.

For directed graphs two different nodes are said to belong to the same connected class if there are paths in both the directions between them, i.e. if they are contained in a circuit path. Thus the nodes i and j ($i \neq j$) are in the same connected class if $j \in \hat{G}(i)$ and $i \in \hat{G}(j)$, i.e. if $\hat{A}_{ij} \hat{A}_{ji} = 1$. The nodes in a connected class and the arcs occurring between them constitute a connected component of the graph. Besides these connected components the graph consists of the arcs occurring between nodes in different connected classes. These arcs are characterized by the fact that they are not contained in any circuit path. A directed graph is called connected if it consists of only one connected component.

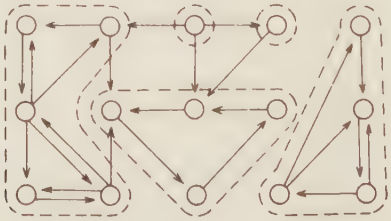


Figure 5

The concepts are illustrated by the graph in Figure 5, which contains five connected components and six arcs between the connected components.

2.4 A connectedness hierarchy

In certain applications it may be of interest to study also connections other than paths, i.e. chains of equally directed arcs.

One may for instance consider chains of arcs which need not be equally directed. This implies that the directed graph is considered as the undirected graph which is obtained when all the arc directions are omitted. If the directed graph has the arc indicator matrix A, the undirected graph has an arc indicator matrix with elements $\max \{A_{ij}, A_{ji}\}$. The connected classes of the undirected graph are called the weakly connected classes of the directed graph.

One may also consider chains of opposite directed pairs of arcs. Then the directed graph is replaced by an undirected graph which is obtained by putting undirected arcs between the nodes which are connected by directed arcs in both the directions in the directed graph. Thus the arc indicator matrix A is replaced by a matrix with elements $\min \{A_{ij}, A_{ji}\}$. The connected classes of the undirected graph are called the strongly connected classes of the directed graph.

The weakly connected, connected, and strongly connected classes of a directed graph constitute a hierachical division of the node set. The node set consists of disjoint weakly connected classes. Each weakly connected class consists of disjoint connected classes. Each connected class consists of disjoint strongly connected classes.

The graph in Figure 5 has two weakly connected classes shown in Figure 6 and eleven strongly connected classes shown in Figure 7.

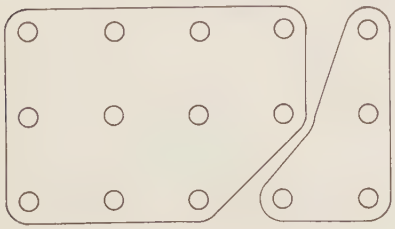


Figure 6

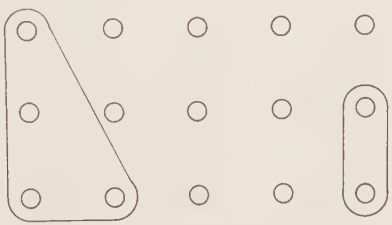


Figure 7

3. DISTANCE

3.1 Definition and simple properties

The distance D_{ij} from node i to node j is defined as the least number of arcs contained in a path from node i to node j. Particularly, D_{ii} is the least number of arcs contained in a circuit path at node i. If there is no path from node i to node j, the distance is infinite, i.e. $D_{ij} = \infty$.

The analysis of the connectedness of the nodes in a graph can be based upon the distance matrix D with the elements D_{ij} for $i \in \Omega$ and $j \in \Omega$. In fact, the path indicator matrix $\hat{A}$ is determined from the distance matrix D according to

$$\hat{A}_{ij} = \begin{cases} 1 & \text{if } D_{ij} < \infty, \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

A graph is connected if, and only if, all the elements of D are finite.

The more detailed information about the graph that is provided by the distance matrix D can be used, except for analysis of connectedness, also for various kinds of centrality analysis which will be demonstrated below in Subsection 3.3.

Some simple properties of the distance matrix D are immediately noted. The possible values of D_{ii} are $1, 2, \ldots, N, \infty$, and the possible values of D_{ij} for $i \neq j$ are $1, 2, \ldots, N-1, \infty$. The inequality

$$D_{ij} \leq D_{ik} + D_{kj} \tag{22}$$

holds for all i, j, and k. If $D_{ij} = s$ and $2 \leq s \leq N$, there exists at least one node k such that $D_{ik} = s-1$ and at least one node l such that $D_{lj} = s-1$. As a consequence the occurrence of a value s in D implies that the value s-1 occurs at least once in the same row as s and at least once in the same column as s for $2 \leq s \leq N$. Particularly, $D_{ii} = N$ implies that $(D_{i1}, D_{i2}, \ldots, D_{iN})$ and $(D_{1i}, D_{2i}, \ldots, D_{Ni})$ are two permutations of $(1, 2, \ldots, N)$.

3.2 Determination of the distance matrix

In order to determine the distance matrix D, use can be made of the Boolean powers of the arc indicator matrix A. This is achieved by noting that D_{ij} is the least value $s \leq N$ such that $(A_s)_{ij} = 1$. If there is no such value of s, the distance is put equal to infinity. The modified indicator matrix B defined by formula (15) can also be used to find D_{ij} for $i \neq j$. In fact, D_{ij} for $i \neq j$ is the least value $s \leq N-1$ such that $(B_s)_{ij} = 1$. If there is no such value of s, the distance is put equal to infinity.

A practical algorithm to determine the distance matrix D may be based upon truncated distances. Introduce for each value of $s = 1, 2, \ldots, N$ a matrix D_s with the elements

$$(D_s)_{ij} = \begin{cases} D_{ij} \text{ if } D_{ij} \leq s, \\ \infty \text{ otherwise.} \end{cases} \tag{23}$$

Consequently,

$$(D_1)_{ij} \geq (D_2)_{ij} \geq \ldots \geq (D_N)_{ij} = D_{ij}, \tag{24}$$

and if $D_s = D_{s-1}$ for a value $s = 2, 3, \ldots, N$, then $D_s = D$. Thus the distance matrix D is obtained as the first repeated matrix in the sequence $D_1, D_2, \ldots$. The matrices D_s of the truncated distances can be successively obtained by application of the following algorithm. The matrix D_1 is given by

$$(D_1)_{ij} = \begin{cases} 1 \text{ if } A_{ij} = 1, \\ \infty \text{ otherwise,} \end{cases} \tag{25}$$

and then

$$(D_s)_{ij} = \begin{cases} (D_{s-1})_{ij} \text{ if } (D_{s-1})_{ij} < \infty, \\ \min\limits_{1 \leq k \leq N} [(D_{s-1})_{ik} + (D_1)_{kj}] \text{ otherwise,} \end{cases} \tag{26}$$

for $s=2,3,\ldots,N$.

If the algorithm is applied manually it is convenient to write A with the zero digits omitted, i.e. D_1 with the infinity marks omitted. In this scheme the empty places are successively filled up. The digit s is put into the place (i,j) if there is a k such that place (i,k) is marked s-1 and place (k,j) is marked 1.

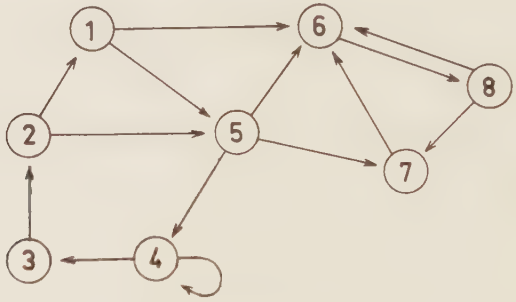


Figure 8

The graph in Figure 8 has the following distance matrix, with infinity marks omitted.

5	4	3	2	1	1	2	2
1	4	3	2	1	2	2	3
2	1	4	3	2	3	3	4
3	2	1	1	3	4	4	5
4	3	2	1	4	1	1	2
					2	2	1
					1	3	2
					1	1	2

3.3 Centrality

Various centrality properties of the graph can be investigated by use of the distance matrix. The most common centrality parameters are the out-radius ρ' and the in-radius ρ'', which are defined by

$$\rho' = \min_i \max_j D_{ij}, \tag{27}$$

and

$$\rho'' = \min_j \max_i D_{ij}. \tag{28}$$

A node i is called an out-center of the graph if it satisfies

$$\max_j D_{ij} = \rho' < \infty, \tag{29}$$

i.e. if each node of the graph has a distance from node i that is at most equal to the finite out-radius. A node j is called an in-center of the graph if it satisfies

$$\max_i D_{ij} = \rho'' < \infty, \tag{30}$$

i.e. if each node of the graph has a distance to node j that is at most equal to the finite in-radius.

The diameter δ of the graph is the largest distance between two nodes of the graph, i.e.

$$\delta = \max_i \max_j D_{ij}. \tag{31}$$

The diameter of a graph is finite if and only if the graph is connected.

Several other parameters have been suggested in the literature to describe in a simple form the centrality properties of a graph. Bavelas (1950), Leavitt (1951), Beauchamp (1965), and Sabidussi (1966) have suggested centrality indices which are based on the minimum value or the arithmetical or harmonical mean value of the row sums of the distance matrix. The use of the harmonical mean values is explained by a desire to have a measure which allows simple treatment of infinite distances.

The row maxima and the row sums of the distance matrix can be interpreted in various fields of application in a way that implies that a certain centrality index is a natural descriptive parameter.

Consider for instance a communication network where the cost of transmission is one unit per message on every arc. If node i transmits different messages to all the nodes which have paths from node i, the cheapest cost of transmission from node i becomes equal to $\sum_j D_{ij}$. It should be noted that the message may be a document, and hence speaking of sending a message to oneself may make sense.

Another example is provided by organization structures represented by graphs without circuit paths. The nodes represent the persons in the organization, and the arcs represent supervision. The number of subordinate nodes to node i becomes equal to the number of finite elements in the i-th row of D. Two nodes may be said to belong to different supervision or status levels if there is a path between them. The least possible number of such levels in the graph can be shown to be equal to the number of nodes contained in the longest path, i.e. $1+\delta$. This and related problems are discussed by Kemeny and Snell (1962).

3.4 Generalization to valued graphs

Instead of measuring distances as the number of arc steps, a more general approach may be applied. Each arc is associated with a value which may be a length, time, cost, etc. By use of the arc values, the path values are defined as the sums of the arc values contained in the paths. The distance D_{ij} is defined as the least value of the paths from node i to node j. D_{ij} may for instance be the length of the shortest path, the time of the quickest path, or the cost of the cheapest path.

It is no essential restriction to assume that the graph has no multiple arcs, because if there are multiple arcs from one node to another, it will suffice to consider only the arc with the least value. Let L_{ij} be a non-negative arc value if there is an arc from node i to node j, and put $L_{ij} = \infty$ otherwise. The distance matrix D can be obtained from the arc value matrix L by application of algorithms that have been given by Ford (1956) and Danzig (1960).

4. CAPACITY

4.1 Introduction

In the analysis of connectedness which was considered in Section 2, the problem was to decide which pairs of nodes were connected by at least one path. In many applications it may be of interest to know not only which nodes are connected but also how many paths there are between the different pairs of nodes. The number of paths which possess certain specified properties may be interpreted as a measure of capacity, and various path properties may be of relevance to different applications.

For instance, a large number of elementary paths in a communication network may correspond to a high degree of reliability against arc disturbances. In transport problems it may be of interest to know the maximum number of paths without a common arc that can possibly be chosen from one node to another. This number corresponds to the transport capacity if every transport unit blocks its path for the other transport units. The maximum number of disjoint paths, i.e. paths without a common arc, also provides a measure of capacity in networks with stationary flows.

4.2 Path frequencies

The methods that were given in Subsection 2.2 to determine the path indicator matrix can easily be modified to provide the path frequency matrix.

Consider the ordinary matrix powers (not the Boolean powers) of the arc frequency matrix C. The element $(C^s)_{ij}$ denotes the number of paths with exactly s arcs from node i to node j. Note that this comprises all the paths, including the non-elementary ones. A simple example will be studied as an illustration. The graph in Figure 9 has the arc frequency matrix

$$C = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

with the square

$$C^2 = \begin{bmatrix} 1 & 1 & 2 & 0 \\ 2 & 0 & 0 & 2 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

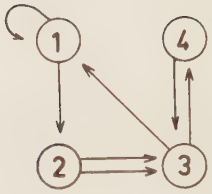


Figure 9

A simple check shows that the number of paths with exactly two arcs is given by the matrix C^2. For instance, $(C^2)_{11} = 1$ corresponds to the path consisting of the circuit arc at node 1 repeated twice.

4.3 Elementary path frequencies

Kaufmann and Malgrange (1963) have developed a method called "latin multiplication" in order to obtain a list of all the elementary paths in graphs without multiple arcs. The method is based on a list of all the paths and a systematical exclusion of the non-elementary paths.

Let $(i_0, i_1, \ldots, i_s)$ be a node sequence contained in an elementary path. If there are multiple arcs, such a node sequence corresponds to

$$\prod_{v=1}^{s} C_{i_{v-1} i_v} \tag{32}$$

different elementary paths. By adding such numbers for all the node sequences representing elementary paths from node i to node j, the total number of elementary paths from node i to node j is obtained. Thus it is no essential restriction that only single arcs should be considered in the "latin multiplication" method.

For low values of s it is possible to apply combinatorial arguments to determine formulas for the number of elementary paths with s arcs. Luce and Perry (1949) give the following expression for the number of elementary paths with three arcs from node i to node j for $i \neq j$:

$$(C^3)_{ij} - (C^2)_{ii}\, C_{ij} - C_{ij}\, (C^2)_{jj} + C_{ij}\, C_{ji}\, C_{ij}. \tag{33}$$

Harary and Ross (1952) give corresponding formulas for four, five, and six arcs.

A method of calculating the number of elementary paths for arbitrary values of s has been suggested by Parthasarathy (1964). He develops an algorithm which is based on the fact that each elementary path from node i to node j $(i \neq j)$ with $s+1$ arcs consists of an arc from node i to some node $k \neq j$ and an elementary path having s arcs from node k to node j which does not contain node i. The algorithm can in a slightly modified version be given in the following way.

First of all it is noted that no circuit arcs can be contained in an elementary path. Consequently, it is no restriction to assume that all the diagonal elements are zero in the arc frequency matrix C, i.e. $C_{ii} = 0$ for $i \in \Omega$. Denote by C_k the arc frequency matrix C with the elements in the k-th column replaced by zeros, i.e.

$$(C_k)_{ij} = \begin{cases} C_{ij} & \text{if } j \neq k, \\ 0 & \text{if } j = k. \end{cases} \tag{34}$$

Moreover, let $f_{sij}(C)$ denote the number of elementary paths from node i to node j with exactly s arcs in the graph with the arc frequency matrix C. Obviously,

$$f_{1ij}(C) = C_{ij} \tag{35}$$

for $i \in \Omega$ and $j \in \Omega$. According to the principle mentioned above, the following recurrence formulas are obtained:

$$f_{s+1,ij}(C) = \sum_{k \neq j} C_{ik}\, f_{skj}(C_i) \tag{36}$$

for $i \neq j$ and $s = 1, 2, \ldots, N-2$,

$$f_{s+1,ii}(C) = \sum_k C_{ik} f_{ski}(C) \tag{37}$$

for $s = 1,2,\ldots,N-1$. Application of (36) and (37) for $s=1$ gives

$$f_{2ij}(C) = \sum_k C_{ik} C_{kj} = (C^2)_{ij} \tag{38}$$

for $i \in \Omega$ and $j \in \Omega$. Application of (36) for $s=2$ gives

$$f_{3ij}(C) = \sum_{k\neq j} C_{ik}(C_i^2)_{kj} = \sum_{k\neq j} C_{ik}\left[(C^2)_{kj} - C_{ki} C_{ij}\right] =$$

$$= (C^3)_{ij} - (C^2)_{ii} C_{ij} - C_{ij}(C^2)_{jj} + C_{ij} C_{ji} C_{ij} \tag{39}$$

for $i\neq j$, which is the same expression as the one given in (33).

4.4 The maximum number of disjoint paths

Two paths are called disjoint if they have no common arc. Note, however, that they may have common nodes. Consider the problem of determining the maximum number of disjoint paths from one node to another, say from node 1 to node N.

Obviously it is no restriction to consider only the elementary paths. Assume that there are N' elementary paths from node 1 to node N, and that it is known which arcs each such path contains. This information can be given in an $R \times N'$ matrix H with elements H_{uv} which are equal to 1 or 0 according to whether arc u is or is not contained in path v. Disjoint paths correspond to orthogonal columns of H, i.e. columns which do not have the digit 1 in the same position. It follows that the maximum number of disjoint paths is equal to the maximum number of orthogonal columns which can be chosen from H. Introduce an undirected graph with N' nodes corresponding to the columns of H and with arcs corresponding to the non-orthogonal pairs of columns. No circuit arcs are attached to the nodes. The maximum number of orthogonal columns of H becomes equal to the maximum number of nodes without any arcs between them.

The original problem is equivalent to the determination of an arc-free node set (Subsection 5.1) with a maximum number of nodes. This problem can be solved by use of the theory of matching according to Berge (1962, Chapter 18).

As a rule, however, it is inconvenient to consider the path frequency problem as a matching problem, since this approach requires that all the elementary paths are listed. In actual fact, it is usual to go the opposite way and consider matching problems as path frequency problems. This is most easily demonstrated by an example.

Suppose that the problem is to choose a maximum number of digits 1 from the matrix

1	1	1	1
0	1	0	0
0	1	0	0

in such a way that at most one digit 1 is chosen from each row and at most one digit 1 is chosen from each column. Introduce a graph with 9 nodes. Node 1 has an arc to each one of the nodes 2, 3, and 4, which correspond to the rows of the matrix. The nodes 5, 6, 7, and 8, which correspond to the columns of the matrix, have arcs to node 9. Between the row and column nodes there are arcs in accordance with the occurrences of digits 1 in the matrix. The graph is shown in Figure 10. Now the matching problem consists of the determination of the maximum number of disjoint paths from node 1 to node 9.

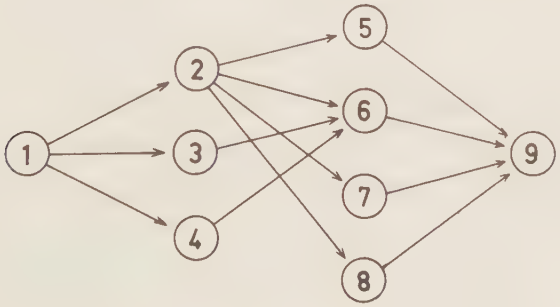


Figure 10

An algorithm to determine the maximum number of disjoint paths from node 1 to node N has been given by Ford and Fulkerson (1962). It is customary to call the arc frequency C_{ij} the capacity from node i to node j, and to search for a maximum flow from node 1 to node N. A flow matrix F has elements F_{ij} which report the amount of flow passing from node i to node j, i.e. the number of the C_{ij} arcs that are occupied by a set of disjoint paths from node 1 to node N. Denote by

$$f_i = \sum_{j=1}^{N} (F_{ij} - F_{ji}) \tag{40}$$

the net outflow from node i for $i \in \Omega$. The conditions a flow matrix has to satisfy can then be given as

$$\left.\begin{aligned} &0 \le F_{ij} \le C_{ij}, \\ &f_1 = -f_N, \\ &f_i = 0 \text{ for } i = 2, 3, \ldots, N-1, \end{aligned}\right\} \tag{41}$$

and the problem is to find the maximum value of f_1. According to a theorem by Ford and Fulkerson this maximum value is equal to the minimum value of

$$\sum_{i \in S} \sum_{j \in \bar{S}} C_{ij} \tag{42}$$

for subsets S of Ω containing node 1 but not node N, i.e. $S \subset \Omega$, $\bar{S} = \Omega - S$, $1 \in S$, and $N \in \bar{S}$.

Boolean programming methods have been applied by Hammer (1965) to determine the minimum value of (42). Introduce variables x_i which are equal to 1 or 0 according to whether $i \in S$ or $i \in \bar{S}$. Then (42) can be written as

$$\sum_{i=1}^{N} \sum_{j=1}^{N} C_{ij}\, x_i(1-x_j), \tag{43}$$

with $x_1 = 1$ and $x_N = 0$. The problem then is to minimize (43) with respect to $x_2, x_3, \ldots, x_{N-1}$. For the graph in Figure 10 the problem is to minimize

$$1-x_2+1-x_3+1-x_4+x_2(1-x_5)+x_2(1-x_6)+x_2(1-x_7)+x_2(1-x_8)+x_3(1-x_6)+$$
$$+x_4(1-x_6)+x_5+x_6+x_7+x_8 = 3+3x_2+(1-x_2)(x_5+x_6+x_7+x_8)-(x_3+x_4)x_6. \qquad (44)$$

The minimum value is 2, and is realized by $x_2 = x_5 = x_7 = x_8 = 0$, $x_3 = x_4 = x_6 = 1$. Several algorithmical methods of minimizing non-linear real valued functions of Boolean variables have been given in the literature. The methods are described in a monography on Boolean methods by Hammer and Rudeanu (1968).

5. INDEPENDENCE AND DOMINANCE

5.1 Independence

An arc-free node set S is a subset of the node set Ω with no arcs between its nodes, i.e. $S \subset \Omega$ and

$$\bigcup_{i \in S} \bigcup_{j \in S} M_{ij} = \emptyset. \qquad (45)$$

The nodes of an arc-free node set are called independent. The maximum number of independent nodes in Ω is denoted α.

Shannon (1956) has used this independence concept to study error-correcting transmission of signals over a noisy channel. The signals are represented by nodes, and the signals that may be confused by the receiver are connected by arcs. For instance, if the signals consist of three binary digits and it is presumed that at most one digit is changed by the transmission, the graph will appear as in Figure 11. The maximum number of signals that can be used in an error-correcting code becomes equal to the maximum number of independent nodes in the graph. In the example $\alpha = 2$, and it is for instance possible to choose the signals 000 and 111.

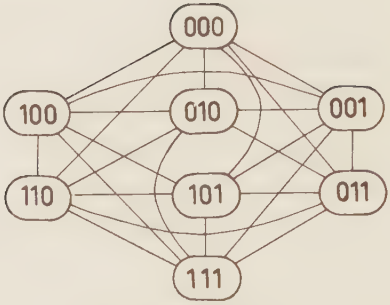


Figure 11

By use of the arc indicator matrix A an arc-free node set S can be characterized by

$$\sum_{i\in S}\ \sum_{j\in S} A_{ij} = 0 . \tag{46}$$

Put x_i equal to 1 or 0 according to whether $i \in S$ or $i \in \bar{S}$. Then the arc-free node sets can be identified with the solutions of the equation

$$\sum_{i=1}^{N}\ \sum_{j=1}^{N} A_{ij}\, x_i\, x_j = 0 . \tag{47}$$

A solution method has been given by Maghout (1959) and is also described by Hammer and Rosenberg (1964) and Hammer and Rudeanu (1968). The maximum number of independent nodes corresponds to the maximum value of $\Sigma\, x_i$ among the solutions of equation (47).

5.2 Chromatic decompositions

A division of the node set Ω into disjoint arc-free subsets is usually called a chromatic decomposition. This term is suggested by the following interpretation. The nodes are colored in such a way that no arc connects two nodes of the same color. The nodes that have the same color constitute an arc-free set.

Problems of map coloring can be stated by use of chromatic decompositions of graphs. The regions of the map are represented by nodes. Arcs are inserted

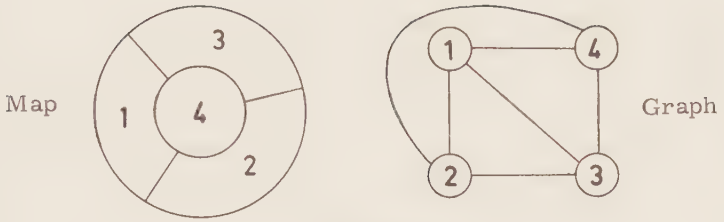


Figure 12

between the nodes that represent regions with a common border line (i.e. those which have more than point contact). A coloring of the map such that no neighboring regions are equally colored corresponds to a chromatic decomposition of the graph.

The Four Color Map Conjecture states that at most four colors are needed for the coloring of a map. Equivalently, every plane graph without circuit arcs has a chromatic decomposition with at most four arc-free sets. A graph is called plane if it can be drawn in one plane in such a way that no arcs intersect outside the nodes. That four colors are needed is seen from Figure 12. A map requiring more than four colors has not been found, and nobody has been able to prove that such a map does not exist. A recent monography on map coloring has been written by Ore (1967).

5.3 Dominance

The nodes in a set S are called dominating if each node outside S has an arc from at least one of the nodes in S, i.e.

$$\bigcup_{i \in S} M_{ij} \neq \emptyset \text{ for } j \in \bar{S}. \tag{48}$$

The minimum number of dominating nodes in Ω is denoted β.

The nodes in a set S are called dominated if each node outside S has an arc to at least one of the nodes in S, i.e.

$$\bigcup_{j \in S} M_{ij} \neq \emptyset \text{ for } i \in \bar{S}. \tag{49}$$

The minimum number of dominated nodes in Ω is denoted γ.

Consider a set of components that constitutes a technical or economical system. The components are exposed to faults or disturbances. Suppose that some sort of control devices are to be placed in some of the components. A control device can detect faults in the component where it is placed and also in some of the other components that are related to it. The system can be represented by a graph where the nodes are the components and the arcs from each node indicate the nodes which can be controlled from it. To control all the nodes, a control device has to be put into each node in a set of dominating nodes. The least number of control devices that are needed to control all the nodes is $\bar{\beta}$.

Consider a graph showing how propaganda, rumour, infection, etc. may circulate among the individuals in a population. The individuals are represented by the nodes and the circulation channels by the arcs. Starting with a set of dominating nodes all the nodes can be reached in the graph. The least possible initial set has to consist of β nodes.

Boolean methods can be applied in order to determine sets of dominating or dominated nodes and β and γ. If x_i is equal to 1 or 0 according to whether $i \in S$ or $i \in \bar{S}$, the condition (48) for S becomes equivalent to the following condition for the Boolean variables. For each $j \in \Omega$ it holds that either $x_j = 1$ or $A_{ij}\, x_i = 1$ for at least one value of $i \in \Omega$. This can be written more concisely by use of the modified arc indicators B_{ij} introduced in formula (15). For each $j \in \Omega$ it holds that $B_{ij}\, x_i = 1$ for at least one value of $i \in \Omega$. Consequently, the condition (48) can be given as

$$\sum_{j=1}^{N} \prod_{i=1}^{N} (1 - B_{ij}\, x_i) = 0 . \tag{50}$$

The condition (49) can analogously be transformed to

$$\sum_{i=1}^{N} \prod_{j=1}^{N} (1 - B_{ij}\, x_j) = 0 . \tag{51}$$

The solution of (50) or (51) and the minimization of $\Sigma\, x_i$ subject to the side conditions (50) or (51) are described for instance by Hammer and Rudeanu (1968).

Chapter 3
STOCHASTIC GRAPHS

SYNOPSIS

1. Introduction. The concept of a stochastic graph is defined. The use of combinatorial methods of computing probabilities is discussed shortly. Three basic means of generating stochastic graphs are described, namely sampling procedures, uncertain measuring procedures, and randomized construction procedures.

2. Sampling models. The need for graph sampling in various applications has been pointed out in the literature, and a few graph sampling procedures have been subjected to a closer examination. A survey is given of nexus sampling, snowball sampling, and social network sampling. In conclusion, some basic sampling and observation procedures in graphs are discussed.

3. Measurement error models. Two applications of measurement error models in graphs are presented. The first one comprises management scheduling problems with uncertain arc times, and the second one comprises flow problems with uncertain arc flows.

4. Randomization models. An account is given of the different randomization procedures which have been described in the literature in order to choose

arcs in a graph. Moreover, a special randomization of nodes is described which has been applied in a model used to study the epidemicity of various deseases. Finally there is a discussion of randomly directed arcs, which are applicable in connection with paired comparisons.

1. INTRODUCTION

1.1 The concept of a stochastic graph

The concept of a stochastic graph can be defined in complete analogy with the concept of a stochastic variable. Starting with a given probability space, a function is defined from the probability space to a set of graphs. Thus the function associates a unique graph with each outcome of the probability space. If the function induces a probability measure on the set of graphs, the function is said to be a stochastic graph. Without having to introduce this explicit consideration of a function, a stochastic graph can be described by a set of graphs together with a probability measure on that set.

1.2 Combinatorial graph problems

Consider a finite set of different graphs and a uniform probability distribution on the set. The probability that a graph randomly chosen from the set will have any particular property is equal to the proportion of graphs in the set with the considered property. In order to compute such probabilities combinatorial methods may be applied to count the number of graphs with a certain property.

In the literature there are several combinatorial studies which provide the numbers of directed and undirected graphs with given node and arc frequencies and with distinguishable and indistinguishable nodes and arcs. Particular attention has been paid in the literature to connected graphs and to connected graphs without circuit paths, referred to as trees. Some references are Pólya (1937), Otter (1948), and Harary (1955). More exhaustive references are provided by Riordan (1967).

In many applications the appropriate set of graphs is not of such a standard composition that the combinatorial methods referred to above can be applied. More often than not, the outcomes of the stochastic graph have to be characterized by a set of graphs utilizing more involved and detailed specifications. In the next subsection such characterizations will be exemplified by three types of generators of stochastic graphs.

1.3 Three basic generators of stochastic graphs

In order to apply stochastic graphs it may be natural to define implicitly the set of possible graphs by describing some device that generates the graphs. It should be noted that it is often possible to find several different devices that generate the same set of graphs. It may be said that the choice of a generating device is part of the model of the practical problem that is to be studied. To facilitate the choice of an appropriate stochastic graph model, it is convenient to have recource to some basic means of generating stochastic graphs.

One such basic means is the application of a random sampling procedure in a given graph. The given graph may be looked upon as a representation of a population or a system of any kind. From this graph the nodes or arcs may be sampled and a portion of the graph observed. The sample graph is a stochastic graph, and its properties are dependent upon the properties of both the population graph and the applied sampling and observation procedures.

Another way of generating a stochastic graph is to introduce errors of measurement in the node or arc variables that are defined in a valued graph. Thus, a certain valued graph is considered which has unknown node or arc values. These values are supposed to be observable only with measurement errors which are appropriately described by stochastic variables. This implies that the observed valued graph is a stochastic graph. If the measurement terminology suggests that the various measurement errors are independent of one another, a more general approach is obtained by allowing the node and arc variables to be arbitrary stochastic variables.

A third way of generating a stochastic graph is the application of a random mechanism to construct a graph. The occurrence of nodes and arcs and the directions of the arcs can be settled randomly. The arcs may be allocated by random devices. The nodes and arcs may be labeled or colored by randomization.

The above three basic means of generating stochastic graphs are not entirely distinct, but can be used in conjunction in particular cases. The purpose of the classification of stochastic graphs in these three categories is mainly to provide

a common basis for the problems that will be taken up in this book. Besides, the classification may have an educational value, since teaching about graph models may benefit by a systematic treatment of stochastic graphs.

In the following sections of this chapter the three basic means of generating stochastic graphs will be more thoroughly discussed and exemplified.

2. SAMPLING MODELS

2.1 Nexus sampling

In a paper by Stephan (1969) modern sample survey theory is discussed, and some fields are suggested for further development of the theory. Among these fields a theory of statistical inference in graphs is contained in what Stephan calls nexus sampling and configurational analysis.

Stephan discusses critically the conventional way of looking at a population as consisting exclusively of units which are associated with measurements or properties characterizing the individual units only. In this way the relationships and interactions that may exist between the population units are completely disregarded. In several connections, however, it may be of value to use such relationships in the population when carrying out a sample survey. Just to provide an example of a simple reason for doing so, it may be imagined that some information about related population units will be accessible only if all these related units are comprised in the sample. Stephan calls attention to the fact that a general theory of nexus sampling, i.e. sampling in populations with a relation structure, is still to be established.

In certain special cases a nexus sampling problem may be turned into a conventional sampling problem by considering the relationships as sampling units. When this is impossible, one has to distinguish between sampling units, observed data, and existing relationships. The relationships may influence both the sampling and observation procedures. The properties of the population that are to be studied may need sampling and observation methods that actually use the existing relationships.

If a population with binary relationships is studied, the population units can

be taken as the nodes and the relationships as the arcs in a graph. Assume that the arcs are associated with arc values, and that the aim of the study is to investigate these arc values. If the arcs can be used as sampling units one obviously has an ordinary sampling problem. But if the nodes have to be used as sampling units, the question arises of which arc values to observe. Generally, the information about the population that is obtained by application of a certain sampling procedure has to be specified. In traditional sampling problems it is customary to assume that values of all the variables which are investigated are observed for each unit in the sample. Occasionally some values may be missing, and then the units which have missing values either have to be omitted or have to be given a complete set of values by extrapolation. When nodes are sampled from a graph, it is important to specify not only the sampling procedure but also the observation procedure. As a concrete illustration, one can imagine that only those arc values are observed which correspond to the arcs between any two nodes in the sample. This implies that the information provided by a node in the sample depends upon the other nodes which are contained in the sample. Generally, a characteristic feature of nexus sampling is the system aspect of the inherent information of a sample, i.e. the information depends upon the whole sample and cannot be broken down into information units associated with the sampling units.

2.2 Snowball sampling

Goodman (1961) has used the term snowball sample for a sample which can be described in the following way if a population of individuals is considered. First an initial sample is selected according to some random sampling procedure. Each one of the individuals in this sample is asked to name his k best friends in the population. The individuals who are not contained in the initial sample but who are named by at least one person in the initial sample are said to be in stage 1. Each individual in stage 1 is asked to name his k best friends in the population. The individuals who are not in the initial sample and not in stage 1 but who are named by at least one person in stage 1 are said to be in

stage 2. Each individual in stage 2 is asked to name his k best friends in the population, etc. An s-stage sample comprises the individuals who are in the initial sample or any of the first s stages.

The probability that an individual will be a member of an s-stage sample depends upon the sampling procedure that is used to choose the initial sample and also the friendship structure of the population. The friendship relations can be given as the arcs in a directed graph, where the nodes represent the individuals in the population. Each node has k out-arcs. No circuit arcs and no multiple arcs are present. If the graph has N nodes, it has R=Nk arcs. An s-stage sample consists of the nodes which are reached in at most s arc steps from one of the nodes in an initial sample.

In principle an s-stage sample can be used to study arbitrary node or arc variables in a valued graph. However, Goodman uses the s-stage sample only to study the friendship relations in the population. He assumes that the out-arcs are known for each node in the s-stage sample, i.e. also for the nodes in stage s. The structure of the friendship relations are described by certain parameters m_{sk}. For fixed values of s and k, the parameter m_{sk} is defined as the number of groups with s+k individuals which have the property that each individual in the group has at most s arc steps to any other member of the group and no arcs to non-members of the group. Particularly, m_{1k} is the number of groups with k+1 individuals which have all their selected friends in the group.

The parameter m_{sk} is to be estimated by use of the information about the k best friends of each one of the individuals in an s-stage sample. The initial sample is supposed to be a Bernoulli sample, i.e. each individual in the population has a probability p of being selected, and the individuals are selected independently of each other. Put q=1-p. Write m instead of m_{sk}. Let $\Omega_1, \Omega_2, \ldots, \Omega_m$ be the m subsets of the node set Ω which all consist of s+k nodes which have at most s arc-steps to any other node in the same subset and no arc going out from the subset. The subsets $\Omega_1, \Omega_2, \ldots, \Omega_m$ are disjoint. Put $\Omega_0 = \Omega - \Omega_1 - \Omega_2 - \ldots - \Omega_m$. Let $n_0, n_1, \ldots, n_m$ denote the number of nodes in $\Omega_0, \Omega_1, \ldots, \Omega_m$, respectively, which belong to the initial sample. Consequently,

$n_0, n_1, \ldots, n_m$ are stochastically independent, n_0 is binomial $(N-m(s+k), p)$ and n_v is binomial $(s+k, p)$ for $v = 1, 2, \ldots, m$.

The subsets among $\Omega_1, \Omega_2, \ldots, \Omega_m$ from which at least one node is selected to the initial sample will be entirely contained in the s-stage sample. It follows that the positive numbers among $n_1, n_2, \ldots, n_m$ will be observed in the s-stage sample. Let w_j denote the number of subsets among $\Omega_1, \Omega_2, \ldots, \Omega_m$ which have exactly j nodes in the initial sample, i.e.

$$w_j = \sum_{v=1}^{m} I(n_v=j) \tag{1}$$

for $j = 0, 1, \ldots, s+k$. Here I(S) denotes an indicator variable that equals 1 or 0 according to whether the statement S is correct or not. Consequently, the variables $w_1, w_2, \ldots, w_{s+k}$ are observable. Since it holds that the sum

$$\sum_{j=1}^{s+k} j w_j = \sum_{v=1}^{m} n_v \tag{2}$$

and the total node frequency $\sum_{v=0}^{m} n_v$ of the initial sample are both observable, it follows that n_0 is observable.

The vector variable $(w_0, w_1, \ldots, w_{s+k})$ is multinomial $(m, P_0, P_1, \ldots, P_{s+k})$ with

$$P_j = \binom{s+k}{j} p^j q^{s+k-j} \tag{3}$$

for $j = 0, 1, \ldots, s+k$. Moreover, $(w_0, w_1, \ldots, w_{s+k})$ and n_0 are stochastically independent. Goodman considers arbitrary (also non-linear) unbiased estimators of m based upon the observable variables $n_0, w_1, w_2, \ldots, w_{s+k}$. He shows that when N is unknown the minimum variance estimator equals

$$\hat{m} = (w_1+w_2+\ldots+w_{s+k})/(P_1+P_2+\ldots+P_{s+k}) = x/(1-P_0), \tag{4}$$

where $x = w_1+w_2+\ldots+w_{s+k}$ is the number of subsets among $\Omega_1, \Omega_2, \ldots, \Omega_m$

which are observed in the s-stage sample. Since x is binomial $(m, 1-P_0)$, it follows that the minimum variance equals

$$\text{Var } \hat{m} = m\, P_0/(1-P_0). \tag{5}$$

Goodman also reports an upper bound of the expected number of persons interviewed, i.e. contained in the s-stage sample. Consider a fixed node j and let u_j be the number of other nodes which have at most s arc steps to node j. Node j is contained in the s-stage sample with the probability $1-q^{1+u_j}$ and the expected number of nodes in the s-stage sample becomes

$$\sum_{j=1}^{N} \left(1-q^{1+u_j}\right) = N\left(1-\sum_{j=1}^{N} q^{1+u_j}/N\right). \tag{6}$$

Since q^u as a function of u is convex from below for positive values of u, it follows that the expression in (6) is at most equal to

$$N\left(1-q^{\Sigma(1+u_j)/N}\right). \tag{7}$$

The sum $\sum_{j=1}^{N} u_j$ is equal to the number of ordered pairs of different nodes which are situated at most s arc steps apart. Consider a fixed node i and its distances to the other nodes. There are k nodes with distance 1 from node i. There are at most k^2 nodes with distance 2 from node i, etc. It follows that

$$\sum_{j=1}^{N} u_j \leq N(k+k^2+\ldots+k^s), \tag{8}$$

and consequently the expression in (7) is at most equal to

$$N\left(1-q^{1+k+k^2+\ldots+k^s}\right), \tag{9}$$

which is the upper bound Goodman reports for the expected number of nodes in the s-stage sample.

2.3 Sampling social networks

Graphs showing communication, supervision, friendship or any other kind of relationship between pairs of individuals, households, communities or any other kind of social unit are called sociograms or social networks in the sociometry literature. The difficult measurement problems that are often involved in establishing the relationships have probably been the main reason why extensive sociogram analysis is rather unusual. Another reason may be the lack of effective sampling methods in graphs and the lack of methods of analysis of stochastic graphs.

Most studies of sociograms which have been described in the literature do not involve any statistical inference to a population network. Some examples of fields of application which have been treated as sociograms are interpersonal influences in market behavior (Katz and Lazarsfeld, 1955), diffusion of drug news among doctors (Coleman, Katz, and Menzel, 1957), farm practice diffusion among neighbouring farms (Katz, 1960), and rumor transmission (Buckner, 1965).

The problems of measurement and sampling of sociograms have been discussed by Coleman (1958) and Proctor (1966 and 1967). Proctor describes some sampling investigations of social relationships and points out practical difficulties which are involved in the collection and processing of the data. Proctor emphasizes the dominating role played by the measuring problems, but also the need of effective sampling and data processing methods.

2.4 Sampling and observation procedures in a graph

Sampling from a graph can be carried out in a variety of different ways. The sampling units can be nodes, pairs of nodes, arcs, paths, flow units, etc. The sampling procedure can be simple random sampling, stratified sampling, cluster sampling, multistage sampling, etc. The sampling procedure may depend upon the graph structure.

In a concrete problem the choice of sampling units can be determined by the very nature of the problem. It is also possible to imagine cases when it is in-

teresting to compare the efficiency of sampling procedures with different kinds of sampling unit.

Together with every sampling procedure one has to specify the way in which the observations are given. The observation procedure can also be determined by the actual problem one studies, and as a rule it is dependent upon the graph structure.

In an earlier paper by the author (Frank, 1969 a), some basic sampling and observation procedures in graphs have been discussed. These procedures will be described here together with some related procedures.

Consider a graph with N nodes, node set $\Omega = \{1, 2, \ldots, N\}$ and arc sets M_{ij} for $i \in \Omega$ and $j \in \Omega$. Assume that the total arc set M has R arcs.

Let $S \subset \Omega$ be a node sample obtained by some random sampling procedure. In many fields of application one has to consider certain basic sample graphs generated by the node sample S. In sampling social networks, for instance, the arcs between the sampled nodes are usually found by interviews. It may also be the case that all the out-arcs or all the in-arcs at the sampled nodes may be obtained by interviews. Formally, the corresponding sample graphs can be given as in the following schedule.

Node sample S and arc registration	The sample graph has	
	Node set	Arc sets
Within S	S	M_{ij} for $(i,j) \in S \times S$
From S	Ω	M_{ij} for $(i,j) \in S \times \Omega$
To S	Ω	M_{ij} for $(i,j) \in \Omega \times S$
From and to S	Ω	M_{ij} for $(i,j) \in S \times \Omega \cup \Omega \times S$

The first case gives a sample graph that is a subgraph of the population graph. This case will be treated in Chapter 4. The other three cases give sample graphs that are partial graphs of the population graph. These cases will be taken up in Chapter 5.

Bloemena (1964) has considered sampling from a graph in the following way. Two disjoint node samples S_1 and S_2 are selected by random sampling. The arcs within S_1, within S_2, and between S_1 and S_2 are registered. By confining the study to only one of the two samples, the same kind of sample graph will be obtained as in the first case above. Bloemena investigates moment properties of sums $\sum_i \sum_j w_{ij}$ of stochastic variables w_{ij} which satisfy the conditions $w_{ij} \equiv w_{ji}$ and $w_{ii} \equiv 0$. The results are applied to the arc frequencies within S_1, within S_2, and between S_1 and S_2. In Chapter 4 analogous results will be derived by use of algebraic methods that are simpler than the methods of Bloemena and suitable for the statistical inference problems. No inference problems are considered by Bloemena. He gives sufficient conditions for the arc frequencies within S_1, within S_2, and between S_1 and S_2 to be approximately normally distributed. He also gives sufficient conditions for the arc frequencies to converge almost certainly to zero when the node frequency N approaches infinity. These results can be used to give sufficient conditions for the corresponding asymptotic properties of the estimators of the arc frequency R. Such problems, however, will not be dealt with in this book.

Consider an arc sample $T \subset M$ obtained by some random sampling procedure. Let T_1 and T_2 be the sets of start and end nodes respectively of the arcs in T. Some examples of sample graphs generated by the arc sample T are given in the following table.

Arc sample T and arc registration	The sample graph has	
	Node set	Arc sets
Within T	$T_1 \cup T_2$	$M_{ij} \cap T$ for $(i,j) \in T_1 \times T_2$
From T_1	Ω	M_{ij} for $(i,j) \in T_1 \times \Omega$
From and to T_1	Ω	M_{ij} for $(i,j) \in T_1 \times \Omega \cup \Omega \times T_1$
From T_1 and from T_2	Ω	M_{ij} for $(i,j) \in (T_1 \cup T_2) \times \Omega$

These examples show that even when the sampling procedure is a rather simple one, there are numerous possibilities of varying the observation procedure.

3. MEASUREMENT ERROR MODELS

3.1 Scheduling problems

In management scheduling problems a project is considered that is composed of several activities which have to be performed in accordance with a prescribed partial order system. In the simplest case each activity is associated with a performance time, and it is required to find the minimum time needed to get the whole project finished.

The project can be represented by a valued graph with the activities given as the arcs. At each node all the in-arcs correspond to activities which have to be finished before any of the activities corresponding to the out-arcs can be started. The arcs are associated with the performance times. The graph has no circuit paths. The graph may have arcs corresponding to dummy activities with zero performance times. Using the dummy activities, it is no restriction to assume that the graph has a start node with only out-arcs and an end node with only in-arcs. Each path from the start node to the end node is associated with a path time equal to the sum of the arc times along the path. The minimum time required for the project corresponds to the maximum path time, and a path from the start node to the end node which realizes the maximum path time is called a critical path.

The performance times of the activities may be uncertain, due to measurements or estimations which cannot be obtained without errors. The uncertainty or measurement errors in the arc times may be appropriately described by random variables. In several works about network planning the arc times have actually been considered as stochastic variables. The required minimum project time then becomes a stochastic variable, and the main problem is to determine its distribution.

Many simplifying assumptions and approximations have been suggested to make this problem easier. The arc times have been replaced by their expected values, and the algorithms which provide the minimum project time can then be applied to give the minimum expected project time. The minimum expected project time is greater than or equal to the expected minimum project time. Hence,

an upper bound of the expected minimum project time is obtained in this way. In order to obtain some knowledge of the variance of the minimum project time the following method has been suggested. Consider a path realizing the minimum expected project time. Take the sum of the variances of the arc times along this critical path. This sum is used as an approximation to the variance of the minimum project time. Several other approximation and simulation methods have also been suggested and analyzed in the literature. Evidently, such methods are motivated by the difficulties involved in the stochastic scheduling problems. Some references of interest in this connection are Kelley (1961), Fulkerson (1962), Van Slyke (1963), MacCrimmon and Ryavec (1964), and Hildenbrand (1966).

3.2 Flow problems

Consider a graph with any kind of flow quantity associated with the arcs. The arc flows may have to satisfy various types of condition. There may be restrictions imposed by lower and upper bounds of the arc flows. There may be conditions on the sums of the inflows and the outflows at the nodes. There may be conditions on the net outflows at the nodes, for instance equalities or inequalities that have to be satisfied by the net outflows.

Suppose that the arc flows in a graph are incompletely known, but that further information may be obtained by some measurement or observation procedure. If stochastic measurement errors are introduced in the observed data, the observed graph can be considered as a stochastic graph. The unknown arc flows are to be estimated by use of the observed data and the known restrictions and conditions on the arc flows. The probability distributions of the measurements have to be specified. For instance, it may be assumed that the measurements have no systematic errors and independent random errors with equal variances. With assumptions of this kind the estimation problem can be formulated as a least squares problem with side conditions. A thorough discussion of theory, methodology, and applications of such problems is given in Chapter 6.

4. RANDOMIZATION MODELS

4.1 Random arcs

A simple way of introducing randomness in a graph is to let the arcs be randomly chosen. This can be done in several ways, and some procedures which have been suggested in the literature will be considered here.

Erdös and Rényi (1959 and 1960) construct undirected stochastic graphs with N nodes and R arcs by placing arcs between R pairs of nodes which are randomly chosen from the set of the $\binom{N}{2}$ possible unordered pairs. Erdös and Rényi consider the graph construction as an evolutionary process by choosing one arc per unit of time. At time 1, one arc is chosen at random from the $\binom{N}{2}$ possible, at time 2 another arc is chosen at random from the $\binom{N}{2}-1$ possible, etc. In this way they obtain a stochastic process with graph states, and they consider the probability that the graph is connected at a certain stage in the process and the probability that the first connected graph occurs before a certain time limit, etc.

In a paper by Austin et al. (1959) the $\binom{N}{2}$ node pairs are sampled R times with replacement, and the R arcs are labeled so that there are $\binom{N}{2}^R$ possible outcomes. The probability distribution of the number of connected components is determined.

An analogous treatment has been applied by Palasti (1966) to directed stochastic graphs. He studies the probability that the graph is connected.

Another way of specifying a stochastic graph has been suggested by Gilbert (1959). He constructs an undirected graph with a stochastic arc frequency by considering the $\binom{N}{2}$ possible pairs of nodes. Between each pair of nodes an arc is inserted with probability p and not inserted with probability $q = 1-p$. The arcs between different pairs of nodes are independent. Gilbert determines the probabilities that two fixed nodes are path-connected and the probability that the graph is connected.

Directed stochastic graphs with a fixed number of out-arcs from each node have been applied and discussed in sociometry. Let $a_1, a_2, \ldots, a_N$ be the prescribed out-degrees of the graph. Circuit arcs and multiple arcs are not allowed.

The a_i out-arcs from node i are allocated at random to the N-1 other nodes. In stochastic graphs of this type some problems which have been considered are, for instance, calculation of the mean value and the variance of the number of reciprocal choices, and determination of the probability distribution of the number of isolated nodes. Some references are Katz (1951), Katz and Powell (1954 and 1957), and Weibull (1958). This type of stochastic graph has also been treated by Bhargava (1965), who studies the mean value and the variance of the number of subgraphs with n nodes which are connected, weakly connected, or disconnected. Problems of this kind have also been considered by Frank (1966), who gives short proofs and some extensions of various results from the sociometry literature.

In a survey paper by Bhargava and Fisk (1967), there is a very superficial discussion of various ways of constructing stochastic graphs by random choice of arcs.

The stochastic graphs constructed according to the principles described here will be systematically analyzed in Chapter 7.

4.2 Random intersection of two graphs

A stochastic graph constructed by randomization of nodes has been applied by Barton and David (1966) to a study of time and space interaction of the incidence of various diseases, i.e. an investigation of whether there is a tendency for cases to form clusters in time and space, which would be evidence for epidemicity.

First, two graphs are constructed with the reported cases as the nodes, say N nodes. In one of the graphs arcs connect the nodes which correspond to cases that occur at about the same time, and in the other graph arcs connect the nodes which correspond to cases that occur at neighboring places. Closeness in time and space have to be precisely defined of course.

The number R of pairs of nodes which are arc-connected in both the time graph and the space graph is used to test whether there is any interaction between closeness in time and closeness in space. If there is no interaction,

the distribution of R is considered to be given by assigning equal probabilities to the N! possible ways of identifying the nodes in the time graph and the space graph. Consequently, R can be considered as the arc frequency in a stochastic graph which is a random intersection of two fixed graphs. Barton and David determine the factorial moments of R and show that R is asymptotically Poisson-distributed, provided that certain conditions are satisfied.

4.3 Random directions

David (1963) has described various models for studying paired comparisons between N objects. The objects can be taken as the nodes in a graph, and the outcomes of the comparisons can be given by directed arcs. In several models the directions of the arcs are considered to be randomly chosen.

In order to treat consistency problems of the preferences given by the paired comparisons, Kendall and Smith (1940) have suggested a criterion based on the number of circuit paths with three nodes. This number is a stochastic variable with a distribution that depends on the probability assumption about the preferences given by the paired comparisons.

A treatment of consistency problems and other problems for certain graphs with randomly directed arcs will be given in Chapter 8.

Chapter 4

INFERENCE FROM SAMPLED SUBGRAPHS

SYNOPSIS

1. The population graph. A notation system is introduced for the frequency distribution of the components of a vector or matrix. General notations of the mean values, the variances, and the covariances are introduced. Notations of the population graph are given. Relationships between various mean values and variances in the population graph are studied.

2. The sample graph. The sampling distribution of subgraphs is discussed, and notations of the sample graph are given. A graph list is provided to illustrate the sampling distributions of subgraphs and of certain characteristics of the sample graph. Sample selection indicator variables are used to study the stochastic properties of the basic arc frequencies of the sample graph.

3. Estimation of the arc frequency. The total number of arcs is estimated for an undirected population graph which has no multiple arcs and no circuit arcs. The variance of the estimator is unbiasedly estimated. The results are illustrated by a numerical example and by an application to sociometry.

4. Estimation of the node and arc totals. A general population graph is considered, with numerical values attached to the nodes and to the arcs. The sums

of the values of the nodes and of the arcs are estimated from a sampled subgraph, which is chosen according to a general sampling scheme. The sampling scheme comprises sampling with and without replacement and Bernoulli sampling. The results clearly show how the estimators are built up partly from characteristics of the sampling scheme and partly from characteristics of the population graph. The corresponding appropriate parametrization of the population graph is discussed. Subgraph sampling from a valued graph may be looked upon as a matrix sampling design. The use of matrix sampling is illustrated by an example.

5. Estimation of the distribution of the local degrees. An estimation method is suggested to find the number of nodes of the population graph that have a fixed number of arcs. The results are rather incomplete. Unbiased estimators are provided, but no variance estimators. The estimation of the number of isolates is considered as a special case.

6. Estimation of the number of connected components. This section deals mainly with the case of complete connected components. An unbiased estimator is given together with its variance. Explicit simple formulas are obtained when Bernoulli sampling is used. When the connected components are trees the estimation problem is shown to be solvable by using the results of Section 4.

1. THE POPULATION GRAPH

1.1 Preliminaries

If a vector a has N components $a_1, a_2, \ldots, a_N$, the number of components that are equal to the value u will be written

$$f_a(u) = \sum_{j=1}^{N} I(a_j=u). \tag{1}$$

Here I(S) denotes an indicator variable that equals 1 or 0 as the statement S is correct or not. For different values of u, the frequencies $f_a(u)$ give the distribution of the values of the components of the vector a. The corresponding relative frequencies are denoted by $p_a(u) = f_a(u)/N$. The mean value of the components of the vector a is written

$$m(a) = \sum_{j=1}^{N} a_j/N, \tag{2}$$

and the variance is written

$$s^2(a) = \sum_{j=1}^{N} [a_j-m(a)]^2/N = \sum_{j=1}^{N} a_j^2/N - m^2(a). \tag{3}$$

For two N x 1 vectors a and b, the simultaneous distribution and the covariance are defined according to

$$f_{a,b}(u,v) = \sum_{j=1}^{N} I(a_j=u,\ b_j=v) = N\, p_{a,b}(u,v), \tag{4}$$

$$s(a,b) = \sum_{j=1}^{N} [a_j-m(a)]\,[b_j-m(b)]/N = \sum_{j=1}^{N} a_j b_j/N - m(a)\, m(b). \tag{5}$$

For an N x N square matrix C it will be useful to consider the N(N-1) off-diagonal elements separately. Define

$$m(C) = \sum_{i \neq j} \sum C_{ij}/N(N-1), \tag{6}$$

and

$$s^2(C) = \sum_{i \neq j} \sum C_{ij}^2/N(N-1) - m^2(C). \tag{7}$$

For a symmetrical $N \times N$ matrix $C=C'$ it may be noted that

$$m(C) = \sum_{i<j} \sum C_{ij}/\binom{N}{2}, \tag{8}$$

and

$$s^2(C) = \sum_{i<j} \sum C_{ij}^2/\binom{N}{2} - m^2(C). \tag{9}$$

For two $N \times N$ matrices A and B, the following notation will be used

$$s(A,B) = \sum_{i \neq j} \sum A_{ij} B_{ij}/N(N-1) - m(A)\, m(B). \tag{10}$$

1.2 The population graph

The notations introduced in Chapter 2, Subsections 1.2 and 1.3, will be used for the population graph. Thus, the population graph has N nodes, node set $\Omega = \{1,2,\ldots,N\}$, and arc sets M_{ij} for $i \in \Omega$ and $j \in \Omega$. The total arc set M has R arcs. The arc frequency matrix is C, and the local out-degrees, in-degrees, and circuit arc frequencies are given by the vectors a, b, and c, respectively.

It follows immediately from the definitions that the mean values of the vectors a and b and the matrix C satisfy the equalities

$$m(a) = m(b) = (N-1)\, m(C). \tag{11}$$

If the graph is directed, it also holds that

$$m(a) + m(c) = m(b) + m(c) = R/N. \tag{12}$$

If the graph is undirected, it holds that a=b and

$$m(a) + 2\, m(c) = 2\, R/N. \tag{13}$$

By standard partitioning of the sum of squares, it follows that

$$N(N-1)\, s^2(C) = \sum_{i \neq j} \sum [C_{ij} - a_i/(N-1)]^2 + N\, s^2(a)/(N-1), \tag{14}$$

and consequently

$$s^2(a) \leq (N-1)^2\, s^2(C),$$

with equality if, and only if, $C_{ij} = a_i/(N-1)$ for $i \neq j$, i.e. the off-diagonal elements are equal in each row of C. In the same way,

$$s^2(b) \leq (N-1)^2\, s^2(C), \tag{16}$$

with equality if, and only if, the off-diagonal elements are equal in each column of C. If the graph has no multiple arcs, it holds that

$$s^2(C) = m(C)\,[1-m(C)], \tag{17}$$

and consequently a directed simple graph has the variance

$$s^2(C) = R\,[N(N-1) - R]/N^2(N-1)^2, \tag{18}$$

and an undirected simple graph has the variance

$$s^2(C) = R\,[\binom{N}{2} - R]/\binom{N}{2}^2. \tag{19}$$

2. THE SAMPLE GRAPH

2.1 The sampling distribution of subgraphs

Consider a population graph with N nodes. From the set of nodes, a random sample of n nodes is chosen without replacement. Each subset with n nodes has the same probability $1/\binom{N}{n}$ of being selected. The subgraph is observed which belongs to the sampled node set. The nodes and also the arcs are regarded as

being indistinguishable, and several node sets may therefore give the same subgraph. The frequency distribution of the different subgraphs that are present among the $\binom{N}{n}$ node sets is called the sampling distribution of the subgraphs with n nodes.

In principle it is possible to determine the sampling distribution of the subgraphs by examining all the $\binom{N}{n}$ samples and distributing them on different subgraphs. When $\binom{N}{n}$ is too large to make such an enumeration tractable, it is possible to approximate the sampling distribution in an empirical way. A great number of independent samples are chosen and distributed on different subgraphs. The absolute frequencies of the empirical distribution obtained are multinomially distributed stochastic variables that in the ordinary way can be used to estimate or to test hypothetical values of the exact sampling distribution of the subgraphs.

2.2 Notations

The numbers of nodes and arcs of the sample graph are denoted n and r respectively. The nodes are labeled $1,2,\ldots,n$ in the order they are chosen. Thus, the same node may have different labels in the population graph and in the sample graph. The arc frequency matrix of the sample graph is an $n \times n$ matrix Z with the elements $Z_{\alpha\beta}$ for $\alpha = 1,2,\ldots,n$ and $\beta = 1,2,\ldots,n$. The out-degree of node α is

$$x_\alpha = \sum_{\beta \neq \alpha} Z_{\alpha\beta}, \tag{20}$$

and the in-degree of node α is

$$y_\alpha = \sum_{\beta \neq \alpha} Z_{\beta\alpha}. \tag{21}$$

Let x, y, and z denote the $n \times 1$ vectors that have as components the local out-degrees, in-degrees, and circuit arc frequencies respectively.

Table 1. Undirected graphs with 2 - 5 nodes and no circuit or multiple arcs.

Graph	Label	Local degrees	Comp. sizes
	2,0	00	11
	2,1	11	2
	3,0	000	111
	3,1	011	12
	3,2	112	3
	3,3	222	3
	4,0	0000	1111
	4,1	0011	112
	4,2,a	1111	22
	4,2,b	0112	13
	4,3,a	1113	4
	4,3,b	1122	4
	4,3,c	0222	13
	4,4,a	2222	4
	4,4,b	1223	4
	4,5	2233	4
	4,6	3333	4

Graph	Label	Local degrees	Comp. sizes
	5,0	00000	11111
	5,1	00011	1112
	5,2,a	01111	122
	5,2,b	00112	113
	5,3,a	00222	113
	5,3,b	01113	14
	5,3,c	11112	23
	5,3,d	01122	14
	5,4,a	11114	5
	5,4,b	11222	23
	5,4,c	02222	14
	5,4,d	01223	14
	5,4,e	11123	5
	5,4,f	11222	5
	5,5,a	11224	5
	5,5,b	12223	5
	5,5,c	11233	5

Table 1. (continued).

Graph	Label	Local degrees	Comp. sizes
	5,5,d	22222	5
	5,5,e	12223	5
	5,5,f	02233	14
	5,6,a	22233	5
	5,6,b	12333	5
	5,6,c	12234	5
	5,6,d	22224	5
	5,6,e	22233	5
	5,6,f	03333	14

Graph	Label	Local degrees	Comp. sizes
	5,7,a	22334	5
	5,7,b	23333	5
	5,7,c	13334	5
	5,7,d	22244	5
	5,8,a	23344	5
	5,8,b	33334	5
	5,9	33444	5
	5,10	44444	5

2.3 A graph list

All the different graphs with n nodes and r arcs that are possible may be shown in a figure for small values of n and r. The number of different graphs increases rapidly with n and r. Formulas for the number of graphs of different kinds, directed or undirected, connected or not connected, with or without circuit arcs, with or without multiple arcs, etc. , may be provided by generating functions obtained by application of the famous theorem of choices by Pólya. Some references are Pólya (1937), Harary (1955 and 1957), and Riordan (1958).

In order to illustrate the sampling distribution of the subgraphs, an appropriate list of all the possible subgraphs of a certain size is needed. In Table 1

the undirected graphs with at most five nodes and without circuit arcs and multiple arcs have been enumerated. The graphs of the list have been labeled with the node frequency, the arc frequency, and a distinguishing letter if there are several graphs with the same node and arc frequencies. The local degrees and the node frequencies of the connected components are given for each graph of the list.

2.4 Characteristics

Some examples of the characteristics of undirected graphs are the total arc frequency, the number of connected components, the number of isolated nodes, and the maximum local degree. The sampling distribution of an arbitrary characteristic is determined from the sampling distribution of the subgraphs. The frequency that pertains to a given value of the characteristic is obtained as the sum of the frequencies that pertain to the subgraphs with the same value of the characteristic.

Usually there are several subgraphs with the same value of the characteristic. In certain simple cases, however, a few characteristics may be enough to determine the subgraph uniquely. For instance, it follows from the graph list above that the graphs with three nodes are uniquely determined by the arc frequency. The graphs with four nodes are identified by the local degrees. Less information will do, however. If the total arc frequency and the number of nodes with only one arc are known, it is possible to determine the graphs with four nodes uniquely. The graphs with five nodes are not identified by the local degrees. In fact, the graphs labeled 5,4,b and 5,4,f have the same local degrees 11222.

2.5 Example 1

Consider the graph of Figure 1 as a population graph. The node frequency is N=8, and the arc frequency is R=12. The graph is undirected and connected. The local degrees are given by a vector a with four components equal to 2 and four components equal to 4, i.e. $f_a(2) = f_a(4) = 4$. The local degrees have the mean

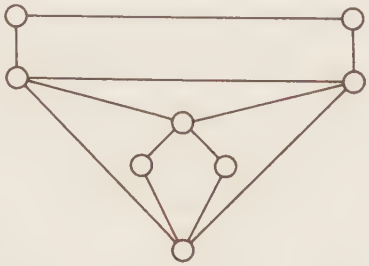


Figure 1

Table 2. Sampling distribution of the subgraphs with 2 - 5 nodes, obtained from the population graph of Figure 1.

Subgraph *	Frequency
2,0	16
2,1	12
3,1	22
3,2	22
3,0	10
3,3	2
4,2,b	26
4,3,b	12
4,4,b	8
4,3,a	6
4,4,a	6
4,1	5
4,2,a	4
4,0	2
4,5	1
5,4,c	8
5,4,f	8
5,5,c	8
5,4,e	6
5,2,b	4
5,3,b	4
5,3,c	4
5,5,b	4
5,5,a	2
5,6,a	2
5,6,c	2
5,6,e	2
5,7,b	2

* The notation of subgraphs is the same as that used in Table 1.

value $m(a) = 3$ and the variance $s^2(a) = 1$.

There are $\binom{8}{n}$ samples with n nodes. For n = 2, 3, 4, and 5 the numbers of samples become 28, 56, 70, and 56, respectively. The distributions of these samples on different subgraphs are given in Table 2. The subgraphs are labeled as in Table 1 above, and they are listed according to non-increasing frequencies.

In Tables 3, 4, and 5 some examples are given of sampling distributions of various characteristics. The sampling distributions of the total arc frequency, the number of isolated nodes, and the number of connected components are given for the sample sizes n = 2, 3, 4, and 5.

Table 3. Sampling distribution of the arc frequency in subgraphs sampled from the population graph of Figure 1.

Arc frequency	Sample size			
	2	3	4	5
0	16	10	2	0
1	12	22	5	0
2		22	30	4
3		2	18	8
4			14	22
5			1	14
6			0	6
7				2
8				0
9				0
10				0
	28	56	70	56

Table 4. Sampling distribution of the number of isolated nodes in subgraphs sampled from the population graph of Figure 1.

Isolated node frequency	Sample size 2	3	4	5
0	12	24	37	40
1		22	26	12
2	16		5	4
3		10		0
4			2	
5				0
	28	56	70	56

Table 5. Sampling distribution of the number of connected components in subgraphs sampled from the population graph of Figure 1.

Connected component frequency	Sample size 2	3	4	5
1	12	24	33	36
2	16	22	30	16
3		10	5	4
4			2	0
5				0
	28	56	70	56

2.6 Example 2

The author has participated in a study of cooperation structures in an organization. The study was carried out by graph methods. The members of the organization were represented by nodes, and the various cooperation contacts that were studied were represented by arcs. As several types of contacts, understood in a wide sence, had to be examined, and as repeated investigations were needed in order to determine the variations over time, it was found appropriate to use sampling methods. For certain special types of contact a total investigation was carried out occasionally. The population graph obtained in such a case is shown in Figure 2.

Figure 2 shows that in real applications the graphs obtained may be rather complicated, and it may even be difficult to form an idea of the appearance of the graph. As a rule it is satisfactory to use the arc frequency matrix as a representation of the graph. In order to obtain a good visual conception of the graph the nodes have to be appropriately labeled. The connected components should appear as submatrices around the main diagonal. It may also be preferable to have the nodes labeled according to some centrality principle within the connected components, i.e. in such a way that the arc frequencies cluster around the main diagonal. In practice it may be hard to achieve good labeling of the nodes. There is need of rapid iterative automatic computer methods in this area.

The population graph in Figure 2 has N=118 nodes and R=253 arcs. The graph consists of nine connected components, of which eight are isolated nodes. The local degrees are given by a vector with the following distribution of the values of the components:

u	0	1	2	3	4	5	6	7	8	9	10	11	12	16	19
$f_a(u)$	8	21	15	17	12	9	7	9	5	5	3	4	1	1	1

The local degrees have the mean value m(a) = 253/59 and the variance $s^2(a)$ = = 42604/3481.

The number of samples with n nodes becomes 6903 for n=2, 266 916 for n=3, 7 673 835 for n=4, and 174 963 438 for n=5. With so many samples it may be

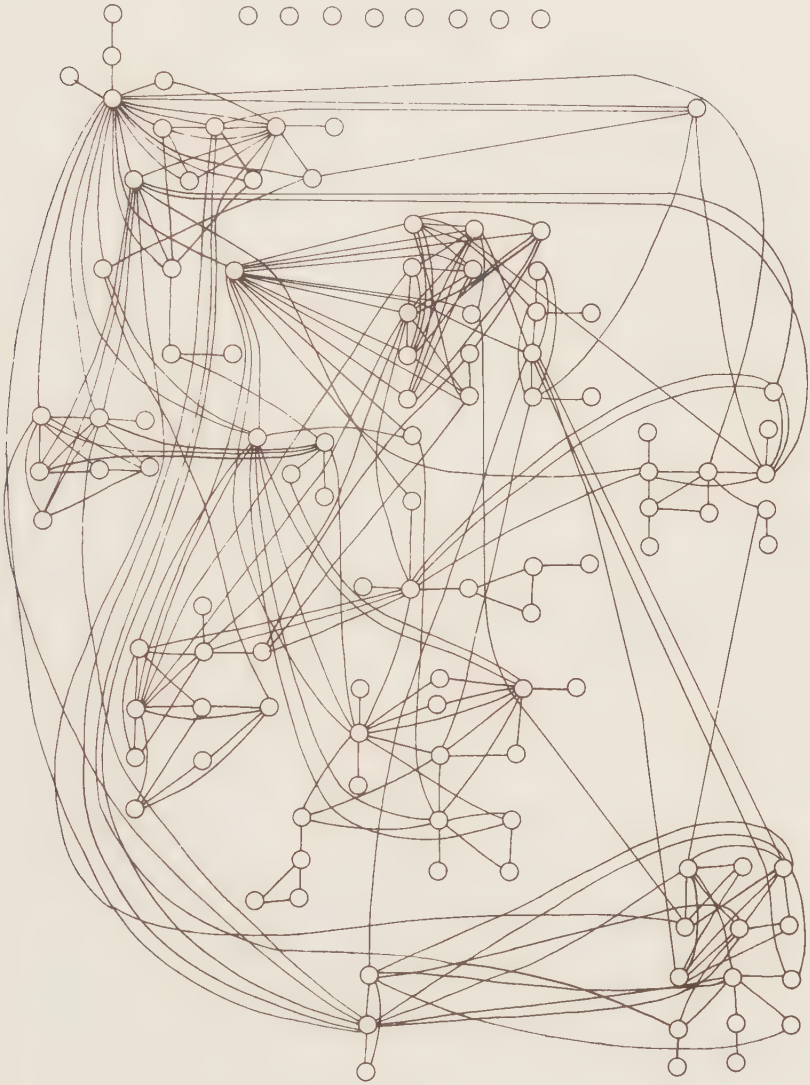


Figure 2

preferable to approximate the sampling distributions by the empirical distributions that are obtained by computer simulation of repeated random sampling of node sets. Tables 6, 7, 8, and 9 show empirical distributions based upon various numbers of samples. Tables 10 and 11 show the corresponding sampling distributions of the arc frequency and the number of connected components.

Table 6. Sampling distribution of the subgraphs with 2 nodes for various numbers of samples chosen from the population graph of Figure 2.

	Frequency %		
	Number of samples		
Subgraph	50	150	250
2,0	94.0	95.3	96.4
2,1	6.0	4.7	3.6
	100.0	100.0	100.0

Table 7. Sampling distribution of the subgraphs with 3 nodes for various numbers of samples chosen from the population graph of Figure 2.

	Frequency %		
	Number of samples		
Subgraph	50	300	500
3,0	88.0	89.0	89.8
3,1	12.0	9.7	9.6
3,2	0.0	1.3	0.6
	100.0	100.0	100.0

Table 8. Sampling distribution of the subgraphs with 4 nodes for various numbers of samples chosen from the population graph of Figure 2.

	Frequency %		
	Number of samples		
Subgraph	100	400	700
4,0	74.0	78.5	79.3
4,1	24.0	19.5	18.7
4,2,a	0.0	0.0	0.1
4,2,b	1.0	1.5	1.4
4,3,a	1.0	0.0	0.0
4,3,b	0.0	0.0	0.3
4,3,c	0.0	0.5	0.1
	100.0	100.0	99.9

Table 9. Sampling distribution of the subgraphs with 5 nodes for various numbers of samples chosen from the population graph of Figure 2.

	Frequency %		
	Number of samples		
Subgraph	200	700	1400
5,0	72.0	66.4	70.2
5,1	23.5	27.3	24.8
5,2,a	2.0	1.7	1.7
5,2,b	2.0	3.6	2.4
5,3,a	0.5	0.6	0.5
5,3,b	0.0	0.3	0.1
5,3,d	0.0	0.0	0.4
5,4,b	0.0	0.1	0.0
	100.0	100.0	100.1

Table 10. Sampling distribution of the arc frequency in subgraphs sampled from the population graph of Figure 2.

	Sample size											
	2			3			4			5		
Arc	Number of samples											
frequency	50	150	250	50	300	500	100	400	700	200	700	1400
0	94.0	95.3	96.4	88.0	89.0	89.8	74.0	78.5	79.3	72.0	66.4	70.2
1	6.0	4.7	3.6	12.0	9.7	9.6	24.0	19.5	18.7	23.5	27.3	24.8
2				0.0	1.3	0.6	1.0	1.5	1.5	4.0	5.3	4.1
3				0.0	0.0	0.0	1.0	0.5	0.4	0.5	0.9	1.0
4							0.0	0.0	0.0	0.0	0.1	0.0
5							0.0	0.0	0.0	0.0	0.0	0.0
6							0.0	0.0	0.0	0.0	0.0	0.0
7										0.0	0.0	0.0
8										0.0	0.0	0.0
9										0.0	0.0	0.0
10										0.0	0.0	0.0
	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.9	100.0	100.0	100.1

Table 11. Sampling distribution of the number of connected components in subgraphs sampled from the population graph of Figure 2.

Connected component frequency	Sample size											
	2			3			4			5		
	Number of samples											
	50	150	250	50	300	500	100	400	700	200	700	1400
1	6.0	4.7	3.6	0.0	1.3	0.6	1.0	0.0	0.3	0.0	0.0	0.0
2	94.0	95.3	96.4	12.0	9.7	9.6	1.0	2.0	1.6	0.0	0.4	0.5
3				88.0	89.0	89.8	24.0	19.5	18.7	4.5	5.9	4.6
4							74.0	78.5	79.3	23.5	27.3	24.8
5										72.0	66.4	70.2
	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.9	100.0	100.0	100.1

2.7 The sample selection indicators

To study the properties of the sampling distributions of various characteristics, it is possible to start with the stochastic properties of the arc frequency matrix of the sample graph. These stochastic properties are to be described by the use of sample selection indicator variables, which will first be defined.

Suppose that n nodes are chosen randomly without replacement from the N nodes of the population graph. Define a sample selection indicator variable $\Theta_{i\alpha}$ equal to 1 or 0 as node i is or is not selected at the α-th draw for $i = 1,2,\ldots,N$ and $\alpha = 1,2,\ldots,n$. The matrix Θ with the elements $\Theta_{i\alpha}$ has $N(N-1)\ldots(N-n+1)$ equally possible outcomes, viz. the matrices that are composed by 0 and 1 in such a way that there is at most one digit 1 in each row and exactly one digit 1 in each column. Thus, $\sum_{\alpha} \Theta_{i\alpha} \leq 1$ for $i=1,2,\ldots,N$, and $\sum_{i} \Theta_{i\alpha} = 1$ for $\alpha = 1,2,\ldots,n$. It follows that

$$
\begin{aligned}
&E\,\Theta_{i\alpha} = 1/N,\\
&E\,\Theta_{i\alpha}\Theta_{j\beta} = 1/N(N-1),\\
&E\,\Theta_{i\alpha}\Theta_{j\beta}\Theta_{k\gamma} = 1/N(N-1)(N-2),\\
&\text{etc.,}
\end{aligned}
\tag{22}
$$

when i, j, k, ... are different numbers chosen from $1,2,\ldots,N$, and when α, β, γ, ... are different numbers chosen from $1,2,\ldots,n$. Put $\theta_i = \sum_{\alpha} \Theta_{i\alpha}$ for $i = 1,2,\ldots,N$. The variable θ_i is equal to 1 or 0 as node i belongs to or does not belong to the sample. From (22) it follows that

$$
\begin{aligned}
&E\,\theta_i = n/N,\\
&E\,\theta_i\theta_j = n(n-1)/N(N-1),\\
&\text{etc.,}
\end{aligned}
\tag{23}
$$

for different numbers i,j,... chosen from $1,2,\ldots,N$.

2.8 Stochastic properties of the arc frequencies

If the nodes of the sample graph are labeled $1,2,\ldots,n$ in the order they are selected, it is possible to write the arc frequency from node α to node β in the

sample graph as

$$Z_{\alpha\beta} = \sum_{i=1}^{N} \sum_{j=1}^{N} C_{ij} \Theta_{i\alpha} \Theta_{j\beta}, \tag{24}$$

for $\alpha = 1,2,\ldots,n$ and $\beta = 1,2,\ldots,n$. In matrix symbols, the arc frequency matrix Z of the sample graph is given as $Z = \Theta' C \Theta$, with the arc frequency matrix C of the population graph and the sample selection indicator matrix Θ. The stochastic properties of Z are then implicitly given by this formula and the stochastic properties of Θ.

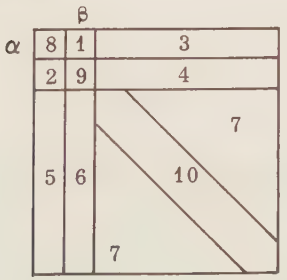
It is an immediate consequence of (24) that

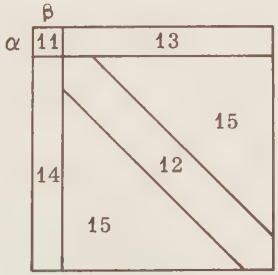
$$E\, Z_{\alpha\beta} = \begin{cases} m(c) \text{ if } \alpha=\beta, \\ m(C) \text{ if } \alpha\neq\beta, \end{cases} \tag{25}$$

for $\alpha = 1,2,\ldots,n$ and $\beta = 1,2,\ldots,n$. To compute the covariance between $Z_{\alpha\beta}$ and $Z_{\gamma\delta}$,

$$E\, Z_{\alpha\beta} Z_{\gamma\delta} = \sum_i \sum_j \sum_k \sum_l C_{ij} C_{kl} E\, \Theta_{i\alpha} \Theta_{j\beta} \Theta_{k\gamma} \Theta_{l\delta} \tag{26}$$

will be considered. Several different cases have to be separated. There are ten cases with $\alpha\neq\beta$ and five cases with $\alpha=\beta$. These cases are illustrated by the schemes





The exact specification of the fifteen cases and the corresponding expressions

of (26) are given by the following list, where all sums are understood to be carried out over different values of the indices. To avoid uninteresting complications, it is supposed that $N \geq 4$.

1. $\sum\sum_{\neq} C_{ij}^2/N(N-1)$ if $\alpha\neq\beta$, $\gamma=\alpha$, $\delta=\beta$,

2. $\sum\sum_{\neq} C_{ij}C_{ji}/N(N-1)$ if $\alpha\neq\beta$, $\gamma=\beta$, $\delta=\alpha$,

3. $\sum\sum\sum_{\neq} C_{ij}C_{ik}/N(N-1)(N-2)$ if $\alpha\neq\beta$, $\gamma=\alpha$, $\alpha\neq\delta\neq\beta$,

4. $\sum\sum\sum_{\neq} C_{ij}C_{jk}/N(N-1)(N-2)$ if $\alpha\neq\beta$, $\gamma=\beta$, $\alpha\neq\delta\neq\beta$,

5. $\sum\sum\sum_{\neq} C_{ij}C_{ki}/N(N-1)(N-2)$ if $\alpha\neq\beta$, $\alpha\neq\gamma\neq\beta$, $\delta=\alpha$,

6. $\sum\sum\sum_{\neq} C_{ij}C_{kj}/N(N-1)(N-2)$ if $\alpha\neq\beta$, $\alpha\neq\gamma\neq\beta$, $\delta=\beta$,

7. $\sum\sum\sum\sum_{\neq} C_{ij}C_{kl}/N(N-1)(N-2)(N-3)$ if $\alpha\neq\beta$, $\alpha\neq\gamma\neq\beta$, $\alpha\neq\delta\neq\beta$, $\gamma\neq\delta$,

8. $\sum\sum_{\neq} C_{ij}C_{ii}/N(N-1)$ if $\alpha\neq\beta$, $\gamma=\alpha$, $\delta=\alpha$,

9. $\sum\sum_{\neq} C_{ij}C_{jj}/N(N-1)$ if $\alpha\neq\beta$, $\gamma=\beta$, $\delta=\beta$,

10. $\sum\sum\sum_{\neq} C_{ij}C_{kk}/N(N-1)(N-2)$ if $\alpha\neq\beta$, $\alpha\neq\gamma\neq\delta\neq\beta$,

11. $\sum C_{ij}^2/N$ if $\alpha=\beta=\gamma=\delta$,

12. $\sum\sum_{\neq} C_{ii}C_{jj}/N(N-1)$ if $\alpha=\beta\neq\gamma=\delta$,

13. $\sum\sum_{\neq} C_{ii}C_{ij}/N(N-1)$ if $\alpha=\beta=\gamma=\delta$,

14. $\sum\sum_{\neq} C_{ii}C_{ji}/N(N-1)$ if $\alpha=\beta=\delta\neq\gamma$,

15. $\sum\sum\sum_{\neq} C_{ii}C_{jk}/N(N-1)(N-2)$ if $\alpha=\beta$, $\alpha\neq\gamma\neq\delta\neq\alpha$. (27)

The expressions are equal in the cases 4 and 5, 8 and 13, 9 and 14, 10 and 15. By straightforward but lengthy algebraic calculations it follows from (27) that the covariance has the following appearance:

$$\mathrm{Cov}(Z_{\alpha\beta}, Z_{\gamma\delta}) = \begin{cases} s^2(C) \text{ in case 1,} \\ s(C,C') \text{ in case 2,} \\ [s^2(a)-(N-1)s^2(C)]/(N-1)(N-2) \text{ in case 3,} \\ [s(a,b)-(N-1)s(C,C')]/(N-1)(N-2) \text{ in cases 4 and 5,} \\ [s^2(b)-(N-1)s^2(C')]/(N-1)(N-2) \text{ in case 6,} \\ [(N-1)s^2(C+C')-2s^2(a+b)]/2(N-1)(N-2)(N-3) \text{ in case 7,} \\ s(a,c)/(N-1) \text{ in cases 8 and 13,} \\ s(b,c)/(N-1) \text{ in cases 9 and 14,} \\ -s(a+b,c)/(N-1)(N-2) \text{ in cases 10 and 15,} \\ s^2(c) \text{ in case 11,} \\ -s^2(c)/(N-1) \text{ in case 12.} \end{cases} \quad (28)$$

If $n \geq 4$, all the cases are present. If $n=3$, case 7 has to be left out. If $n=2$, cases 3, 4, 5, 6, 7, 10, and 15 have to be left out. If the population graph has no circuit arc, the covariance becomes zero in cases $8, 9, \ldots, 15$. If the population graph is undirected, (28) is simplified to

$$\mathrm{Cov}(Z_{\alpha\beta}, Z_{\gamma\delta}) = \begin{cases} s^2(C) \text{ in cases 1 and 2,} \\ [s^2(a)-(N-1)s^2(C)]/(N-1)(N-2) \text{ in cases 3, 4, 5, and 6,} \\ 2[(N-1)s^2(C)-2s^2(a)]/(N-1)(N-2)(N-3) \text{ in case 7,} \\ s(a,c)/(N-1) \text{ in cases 8, 9, 13, and 14,} \\ -2s(a,c)/(N-1)(N-2) \text{ in cases 10 and 15,} \\ s^2(c) \text{ in case 11,} \\ -s^2(c)/(N-1) \text{ in case 12.} \end{cases} \quad (29)$$

3. ESTIMATION OF THE ARC FREQUENCY

3.1 Introduction

When the total arc frequency of a population graph is to be estimated on the basis of a sampled subgraph, it is immaterial whether the arcs are directed or not. If the population graph is directed and has the arc frequency matrix B, it is no real restriction to deal with the corresponding undirected graph that

has the arc frequency matrix C given by

$$C_{ij} = \begin{cases} B_{ij} & \text{if } i=j, \\ B_{ij}+B_{ji} & \text{if } i\neq j. \end{cases} \tag{30}$$

If the directed graph has no circuit arcs, no multiple arcs, and no opposite directed arcs between any pair of nodes, the undirected graph will obviously be without circuit arcs and without multiple arcs, i.e. a simple graph.

If the purpose is to estimate the number of pairs of nodes that have at least one arc in a graph with the arc frequency matrix B, it is equivalent to estimate the arc frequency in the undirected graph with the arc frequency matrix C given by

$$C_{ij} = \begin{cases} 1 & \text{if } B_{ij}+B_{ji} > 0 \text{ and } i\neq j, \\ 0 & \text{otherwise.} \end{cases} \tag{31}$$

Thus one has to deal with an undirected simple graph.

If a directed population graph has the arc frequency matrix B, and if it is required to estimate the number of pairs of nodes that have arcs in both the directions, it is equivalent to consider an undirected graph with the arc frequency matrix C given by

$$C_{ij} = \begin{cases} 1 & \text{if } B_{ij}\, B_{ji} > 0 \text{ and } i\neq j, \\ 0 & \text{otherwise.} \end{cases} \tag{32}$$

The estimation problem is then equivalent to the estimation of the arc frequency in the undirected simple graph.

In the present section an estimation will be made of the total arc frequency of an undirected simple graph. The sampled subgraph will be obtained by node sampling without replacement. The more general type of population graph, i.e.

a directed graph with circuit arcs, multiple arcs, and arc values allowed, will be considered in Section 4, where the sampling scheme will also be generalized.

3.2 An unbiased estimator and its variance

The population graph is undirected and has no circuit arcs and no multiple arcs. The total arc frequency R is to be estimated.

According to (24) the total arc frequency r of the sample graph can be written as

$$r = \sum_{\alpha}\sum_{\beta} Z_{\alpha\beta}/2 = \sum_{i}\sum_{j} C_{ij}\,\theta_i\,\theta_j/2, \tag{33}$$

and it follows from (23) that $E\,r = n(n-1)R/N(N-1)$. Hence an unbiased estimator of R is given by

$$\hat{R} = N(N-1)\,r/n(n-1). \tag{34}$$

To calculate the variance $\sigma^2 = \operatorname{Var}\hat{R}$, the variance

$$\operatorname{Var} r = \sum_{\alpha}\sum_{\beta}\sum_{\gamma}\sum_{\delta} \operatorname{Cov}(Z_{\alpha\beta}, Z_{\gamma\delta})/4 \tag{35}$$

is considered first. The sum in (35) consists of n^4 terms, which can be separated into the fifteen cases defined in (27). There are n terms of case 11. There are n(n-1) terms of each one of the cases 1, 2, 8, 9, 12, 13, and 14. There are n(n-1)(n-2) terms of each one of the cases 3, 4, 5, 6, 10, and 15. There are n(n-1)(n-2)(n-3) terms of case 7. All the terms of the cases $8, 9, \ldots, 15$ are zero since circuit arcs are not present. It follows from (29) that (35) is a linear combination of $s^2(a)$ and $s^2(C)$. By straightforward calculation

$$\begin{aligned}\operatorname{Var} r = {} & (N-n)n(n-1)(n-2)\,s^2(a)/(N-1)(N-2)(N-3) + \\ & + (N-n)(N-n-1)n(n-1)s^2(C)/2(N-2)(N-3).\end{aligned} \tag{36}$$

Hence from (34) it follows that

$$\sigma^2 = \text{Var}\,\hat{R} = N^2(N-1)(N-n)(n-2)s^2(a)/(N-2)(N-3)n(n-1) + \\ + N^2(N-1)^2(N-n)(N-n-1)s^2(C)/2(N-2)(N-3)n(n-1). \quad (37)$$

3.3 An unbiased estimator of the variance

According to the results above, the arc frequency R is unbiasedly estimated by $\hat{R}$. The estimator $\hat{R}$ has a variance σ^2 that depends upon the two population parameters $s^2(a)$ and $s^2(C)$. In order to obtain an unbiased estimator of σ^2 based upon the information provided by the sampled subgraph the corresponding parameters $s^2(x)$ and $s^2(Z)$ of the sample graph will be considered. By writing

$$s^2(x) = \sum_{\alpha} [x_\alpha - m(a)]^2/n - [m(x)-m(a)]^2 \quad (38)$$

and

$$s^2(Z) = \sum_{\alpha \neq \beta}\sum [Z_{\alpha\beta} - m(C)]^2/n(n-1) - [m(Z)-m(C)]^2, \quad (39)$$

and taking expectations, it follows that

$$E\, s^2(x) = \sum_{\alpha} (\text{Var}\, x_\alpha)/n - \text{Var}\, m(x) \quad (40)$$

and

$$E\, s^2(Z) = \sum_{\alpha \neq \beta}\sum (\text{Var}\, Z_{\alpha\beta})/n(n-1) - \text{Var}\, m(Z). \quad (41)$$

As $m(x) = (n-1)m(Z) = 2r/n$ it is possible to obtain Var m(x) and Var m(Z) from (36). Moreover,

$$\text{Var}\, x_\alpha = \sum_{\beta}\sum_{\delta} \text{Cov}(Z_{\alpha\beta}, Z_{\alpha\delta}), \quad (42)$$

and by substitution of (29) and simplification it follows that

$$\text{Var}\, x_\alpha = [(n-1)(n-2)s^2(a) + (N-1)(N-n)(n-1)s^2(C)]/(N-1)(N-2). \quad (43)$$

Substitution of (29), (36), and (43) in (40) and (41) gives after simple but lengthy

calculations

$$\left.\begin{aligned} E\, s^2(x) &= L_{11}\, s^2(a) + L_{12}\, s^2(C), \\ E\, s^2(Z) &= L_{21}\, s^2(a) + L_{22}\, s^2(C), \end{aligned}\right\} \tag{44}$$

with

$$\left.\begin{aligned} L_{11} &= (n-1)(n-2)\,[1-4(N-n)/(N-3)n]/(N-1)(N-2), \\ L_{12} &= (N-1)(N-n)(n-1)(n-2)/(N-2)(N-3)n, \\ L_{21} &= -4(N-n)(n-2)/(N-1)(N-2)(N-3)n(n-1), \\ L_{22} &= 1-2(N-n)(N-n-1)/(N-2)(N-3)n(n-1). \end{aligned}\right\} \tag{45}$$

When the equation system (44) can be solved for $s^2(a)$ and $s^2(C)$, the solutions will be linear expressions of $E\, s^2(x)$ and $E\, s^2(Z)$. If $s^2(x)$ and $s^2(Z)$ are substituted for these expected values, unbiased estimators $\hat{s}^2(a)$ and $\hat{s}^2(C)$ of $s^2(a)$ and $s^2(C)$ will be obtained. When these estimators are substituted in the linear formula (37), an unbiased estimator $\hat{\sigma}^2$ of σ^2 will be obtained. The lengthy calculations are omitted, and only the result is given. When $n \geq 4$,

$$\begin{aligned} \hat{\sigma}^2 = {} & N(N-1)(N-2)(N-n)s^2(x)/(n-1)(n-2)(n-3) - \\ & - N(N-1)(N-n+1)(N-n)\, s^2(Z)/\,2(n-2)(n-3). \end{aligned} \tag{46}$$

When n equals 2 or 3, the equation system (44) is singular, and no estimator $\hat{\sigma}^2$ is obtained.

The estimator $\hat{\sigma}^2$ given by (46) is an unbiased estimator of a positive parameter σ^2, but it may take on negative values. A negative $\hat{\sigma}^2$ is obtained when

$$2(N-2)s^2(x) < (n-1)(N-n+1)s^2(Z). \tag{47}$$

In the next subsection an example is given where both positive and negative values of $\hat{\sigma}^2$ occur. Some further material about the negative variance estimators is given in Subsection 4.7.

3.4 A numerical example

Consider the example of Subsection 2.5 above. The population graph is undirected and has no circuit arcs and no multiple arcs. There are N=8 nodes and R=12 arcs. Moreover, m(a) = 3 and $s^2(a) = 1$. The variance $s^2(C) = R\,[\binom{N}{2}-R]/\binom{N}{2}^2$ becomes equal to 12/49.

Choose samples with n=4 nodes. According to (37) it follows that $\sigma^2 =$ = 1024/45 when n=4. There are 70 possible samples, and the sampling distribution of the subgraphs is given in Table 2. The estimators $\hat{R}$ and $\hat{\sigma}^2$ are given in Table 12. From Table 12 it is easily verified that the estimators are unbiased. It is seen that $\hat{\sigma}^2$ is negative with probability 11/35, zero with probability 1/35, and positive with probability 23/35.

Table 12. The estimators $\hat{R}$ and $\hat{\sigma}^2$ corresponding to the sample size 4 and the population graph shown in Figure 1.

Subgraph	Frequency	$3\hat{R}/14$	$9\hat{\sigma}^2/14$
4,2,b	26	2	32
4,3,b	12	3	-9
4,4,b	8	4	32
4,3,a	6	3	63
4,4,a	6	4	-40
4,1	5	1	11
4,2,a	4	2	-40
4,0	2	0	0
4,5	1	5	11
	70		

3.5 Application to a problem in sociometry

A problem in sociometry that has been paid attention to is to estimate the number of mutual choices between the individuals in a group. The choices may stand for "likes", "has contact with", and so forth. The individuals can be represented by the nodes, and the choices can be represented by the arcs of a directed graph. The parameter to be estimated is the number of pairs of nodes that have arcs in both the directions.

The ordinary approach to the problem is the following. There is a group of N individuals. Each individual makes a fixed number of choices, say k choices. How many reciprocal choices are to be expected when the choices are given at random and independently for the N individuals? With that formulation it can be said that there is a directed graph with N nodes and with stochastic arcs. The problem is to give, for instance, the expected value of the number of mutual choices.

In Chapter 7 a systematic account of such problems will be given. Here a sampling variant will be considered that does not seem to have been studied in any detail in the literature. In two mimeographed papers by Proctor (1966 and 1967) sampling from graphs has been discussed and its importance for sociometric applications has been stressed.

Suppose that a group consists of N individuals and that a sample of n individuals is chosen without replacement. By interviewing the selected individuals information is collected about other individuals in the group which are chosen by each individual in the sample. This information provides the number of reciprocal choices between the individuals in the sample. If the reciprocal choices are represented by the arcs of an undirected graph, the previous results in Subsections 3.2 and 3.3 are directly applicable to an estimation of the number of reciprocal choices in the group.

It is important and must be stressed that no assumptions at all are needed about the group besides knowledge of its size N. The sample graph provides all information that is used. It may be expected that this generality implies that the estimator of the number of mutual choices can be considerably improved by

use of special assumptions about the choice structure of the group. For instance, it would be interesting to investigate how the a priori information that each individual in the group makes k choices can be used to improve the estimator.

4. ESTIMATION OF THE NODE AND ARC TOTALS

4.1 Introduction

The results of the previous section can be generalized in several ways. Instead of the simple population graphs more general population graphs with arc values will be considered. Instead of sampling without replacement a more general sampling scheme will be introduced.

The estimation problem that will be considered is a direct generalization of the estimation of the total arc frequency. Suppose that the population graph is associated with an arbitrary symmetric $N \times N$ matrix C. The totals

$$T_1 = \sum_{i=1}^{N} C_{ii} \text{ and } T_2 = \sum_{i<j} \sum C_{ij} \tag{48}$$

are to be estimated on the basis of a sampled subgraph.

If C is an arc frequency matrix of an undirected graph, T_1 denotes the total number of circuit arcs, and T_2 denotes the total number of the other arcs. Particularly if there are no circuit arcs and no multiple arcs the case of Section 3 is obtained. If an undirected graph without circuit arcs and without multiple arcs has arcs of different values, for instance different lengths or different costs, the total T_2 denotes the sum of the arc values, i.e. the sum of the arc lengths or the arc costs.

The elements of the main diagonal of the matrix C can be regarded as values associated with the nodes, i.e. with the sampling units. To estimate T_1 on the basis of a node sample is a conventional estimation problem that deserves no attention here. However, it may occur that estimators of T_1 and T_2 are considered simultaneously, and as they are based upon the same sampled subgraph they are usually dependent stochastic variables. This is the motivation of considering T_1.

The population graph can be said to be a complete undirected graph with N nodes and $\binom{N}{2}$ arcs. The nodes are associated with the node values C_{ii} for $i = 1,2,\ldots,N$, and the arcs are associated with the arc values C_{ij} for $i<j$.

The sample graph will be assumed to have been obtained from a random node sample and by observation of the node and arc values of the corresponding subgraph. The sampling scheme will comprise, for instance, sampling with and without replacement and Bernoulli sampling.

The results that are obtained for the general setup of this section will throw light upon the results of the previous section and will give a better understanding of the estimation problem. In spite of the more general assumptions that are made, it will prove possible to deal with the problem in a way that is formally compact and that shows how the estimators depend upon the sampling scheme and the appropriate parameters of the sample graph.

4.2 A general sampling scheme

Consider the set of the N nodes of the population graph. A probability is associated with every node subset. The probabilities will be used as selection probabilities when a subset is chosen at random. The probability that a specified node i is an element of the sampled subset will be assumed to be independent of which node i is being considered. Also the probabilities that two, three, or four specified nodes are included in the sampled subset will be assumed to depend only upon the node frequency and not upon the identities of the nodes.

If θ_i denotes a sample selection indicator variable that equals 1 or 0 as node i belongs or does not belong to the sampled subset the assumptions about the sampling scheme can be given as

$$\begin{aligned} &E\,\theta_i = p_1, \\ &E\,\theta_i\theta_j = p_2, \\ &E\,\theta_i\theta_j\theta_k = p_3, \\ &E\,\theta_i\theta_j\theta_k\theta_l = p_4, \end{aligned} \tag{49}$$

where i,j,k,l are different numbers chosen from $1,2,\ldots,N$. Here p_v is the

probability that v specified nodes are included in the sample, and this probability depends on the v nodes only through their number v.

In the case of sampling n nodes without replacement which was considered in the previous section,

$$p_v = \binom{N-v}{n-v} / \binom{N}{n} = \binom{n}{v} / \binom{N}{v} \tag{50}$$

for $v = 1,2,\ldots,n$. In the case of Bernoulli sampling with selection probability p,

$$p_v = p^v \tag{51}$$

for $v = 1,2,\ldots,N$. If m nodes are sampled with replacement, then

$$p_v = \sum_{k=0}^{v} (-1)^k \binom{v}{k} (1-k/N)^m, \tag{52}$$

which can be proved by a combinatorial argument. See for instance Feller (1957, Chapter 4 and the problem sections of Chapters 2 and 4).

The node frequency of the sample graph can be written $n = \sum_i \theta_i$, and it follows from (49) that

$$E\, n = N p_1 \tag{53}$$

and

$$\mathrm{Var}\, n = N(p_1 - p_2) + N^2 (p_2 - p_1^2) . \tag{54}$$

Particularly in the case of sampling without replacement, n is a constant. For Bernoulli sampling with selection probability p, the sample size n is binomial (N, p), and (53) and (54) are simplified to $E\, n = Np$ and $\mathrm{Var}\, n = Np(1-p)$.

4.3 Unbiased estimators and their variances

In order to estimate the node and arc totals T_1 and T_2 of the population graph, the node and arc totals t_1 and t_2 of the sample graph are considered. These totals can be written

$$t_1 = \sum_{\alpha=1}^{n} Z_{\alpha\alpha} = \sum_{i=1}^{N} C_{ii}\,\theta_i\,, \tag{55}$$

and

$$t_2 = \sum\sum_{\alpha<\beta} Z_{\alpha\beta} = \sum\sum_{i<j} C_{ij}\,\theta_i\,\theta_j = \sum\sum_{i\neq j} C_{ij}\,\theta_i\,\theta_j/2\,. \tag{56}$$

Their expected values are immediately obtained as $E\,t_1 = T_1\,p_1$ and $E\,t_2 = T_2 p_2$. Thus, unbiased estimators of T_1 and T_2 are given by $\hat{T}_1 = t_1/p_1$ and $\hat{T}_2 = t_2/p_2$. In order to compute the variances $\sigma_1^2 = \mathrm{Var}\,\hat{T}_1$ and $\sigma_2^2 = \mathrm{Var}\,\hat{T}_2$ of the estimators and their covariance $\sigma_{12} = \mathrm{Cov}\,(\hat{T}_1, \hat{T}_2)$, the expected values $E\,t_1^2$, $E\,t_2^2$, and $E\,t_1 t_2$ are considered first. It follows from (55) that $E\,t_1^2 = T_{11}\,p_1 + T_{12}\,p_2$, with

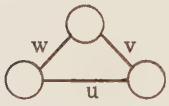
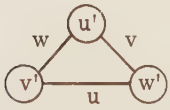
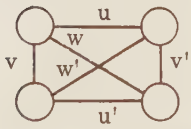
$$\left.\begin{aligned} T_{11} &= \sum_i C_{ii}^2\,, \\ T_{12} &= \sum\sum_{i\neq j} C_{ii}\,C_{jj}\,. \end{aligned}\right\} \tag{57}$$

Here $T_{11} + T_{12} = T_1^2$. It follows from (56) that $E\,t_2^2 = T_{22}\,p_2 + T_{23}\,p_3 + T_{24}\,p_4$, with

$$\left.\begin{aligned} T_{22} &= \sum\sum_{i\neq j} C_{ij}^2/2, \\ T_{23} &= \sum\sum\sum_{\neq} C_{ij}(C_{ik}+C_{jk})/2, \\ T_{24} &= \sum\sum\sum\sum_{\neq} C_{ij}\,C_{kl}/4. \end{aligned}\right\} \tag{58}$$

All the above sums are understood to be carried out over different i, j, k, l. It obtains that $T_{22}+T_{23}+T_{24} = T_2^2$. It follows from (55) and (56) that $E\,t_1 t_2 =$

Table 13. Subgraph totals.

Subgraph with node and/or arc values	Subgraph value	Total over all the subgraphs
	u^2	T_{11}
	2uv	T_{12}
	w^2	T_{22}
	(u+v)w	T_{32}
	2(uv+vw+wu)	T_{23}
	uu'+vv'+ww'	T_{33}
	2(uu'+vv'+ww')	T_{24}

$T_{32}p_2+T_{33}p_3$, with

$$\left.\begin{aligned} T_{32} &= \sum\sum_{i\neq j} C_{ij}\,C_{ii'} \\ T_{33} &= \sum\sum\sum_{\neq} C_{ij}\,C_{kk}/2. \end{aligned}\right\} \qquad (59)$$

The sum defining T_{33} has different i, j, k. Here $T_{32}+T_{33} = T_1T_2$. It follows that

$$\left.\begin{aligned}
\sigma_1^2 &= \operatorname{Var} \hat{T}_1 = T_{11}\,(1/p_1-1) + T_{12}\,(p_2/p_1^2-1),\\
\sigma_2^2 &= \operatorname{Var} \hat{T}_2 = T_{22}\,(1/p_2-1) + T_{23}\,(p_3/p_2^2-1) + T_{24}\,(p_4/p_2^2-1),\\
\sigma_{12} &= \operatorname{Cov}(\hat{T}_1, \hat{T}_2) = T_{32}\,(1/p_1-1) + T_{33}\,(p_3/p_1p_2-1).
\end{aligned}\right\} \qquad (60)$$

4.4 Parametrization of the population graph

The variances and the covariance of the estimators are linear combinations of certain parameters of the population graph with coefficients that depend upon the selection probabilities. The appearing parameters of the population graph can be interpreted as certain subgraph totals and considered as basic graph parameters. To achieve this, the terms in (57), (58), and (59) that belong to the same nodes in different orders are collected and added. By some algebraic calculation if follows that

$$\left.\begin{aligned}
T_{11} &= \sum_i C_{ii}^2,\\
T_{12} &= 2\sum_{i<j}\sum C_{ii}\,C_{jj},\\
T_{22} &= \sum_{i<j}\sum C_{ij}^2,\\
T_{23} &= 2\sum_{i<j<k}\sum\sum (C_{ij}C_{ik}+C_{ij}C_{jk}+C_{ik}C_{jk}),\\
T_{24} &= 2\sum_{i<j<k<l}\sum\sum\sum (C_{ij}C_{kl}+C_{ik}C_{jl}+C_{il}C_{jk}),\\
T_{32} &= \sum_{i<j}\sum C_{ij}(C_{ii}+C_{jj}),\\
T_{33} &= \sum_{i<j<k}\sum\sum (C_{ij}C_{kk}+C_{ik}C_{jj}+C_{jk}C_{ii}).
\end{aligned}\right\} \qquad (61)$$

The parameter T_{11} can be looked upon as a sum over the N subgraphs consisting of just one node. T_{12}, T_{22}, and T_{32} can be looked upon as sums over the $\binom{N}{2}$ subgraphs consisting of two nodes. T_{23} and T_{33} can be looked upon as sums over the $\binom{N}{3}$ subgraphs consisting of three nodes. T_{24} can be looked upon

as a sum over the $\binom{N}{4}$ subgraphs consisting of four nodes. For each subgraph considered, a value is calculated according to Table 13, and then the subgraph totals are obtained by adding the values of all the subgraphs of the same size in the population graph.

Instead of using the subgraph totals as basic parameters of the population graph, it is possible to take certain mean values, variances, and covariances of the node and arc values as basic graph parameters. Then the subgraph totals can be given as functions of these basic parameters. This alternative procedure may seem less natural than the first one, but it has considerable advantages with respect to calculation and interpretation. It is much easier to calculate the ordinary statistical parameters than to make calculations that need enumeration of all the subgraphs with two, three, or four nodes. Moreover, the means and variances are as a rule easier to interpret than the subgraph totals.

The alternative parametrization of the population graph will use the parameters $N, m(a), s^2(a), m(c), s^2(c), s(a,c), m(C)$, and $s^2(C)$ as the basic ones. As $m(a) = (N-1)m(C)$, there are seven basic parameters. By algebraic transformations the subgraph totals are given as the following functions of the basic parameters:

$$\left.\begin{aligned}
T_{11} &= N[s^2(c)+m^2(c)] , \\
T_{12} &= N^2 m^2(c) - N[s^2(c)+m^2(c)] , \\
T_{22} &= \tbinom{N}{2}[s^2(C)+m^2(C)] , \\
T_{23} &= N[s^2(a)+m^2(a)] - 2\tbinom{N}{2}[s^2(C)+m^2(C)] , \\
T_{24} &= \tbinom{N}{2}^2 m^2(C) + \tbinom{N}{2}[s^2(C)+m^2(C)] - N[s^2(a)+m^2(a)] , \\
T_{32} &= N[s(a,c)+m(a)m(c)] , \\
T_{33} &= N\tbinom{N}{2} m(C)m(c) - N[s(a,c)+m(a)m(c)] .
\end{aligned}\right\} \qquad (62)$$

When there are no node values, T_{11}, T_{12}, T_{32}, and T_{33} can be omitted and only T_{22}, T_{23}, and T_{24} remain. Alternatively, $m(c)$, $s^2(c)$, and $s(a,c)$ are omitted, and N, $m(a)$, $s^2(a)$, and $s^2(C)$ remain. If all the arc values are equal to 0 or 1, $s^2(C)$ is redundant, and N, $m(a)$, and $s^2(a)$ remain. For later reference it will be convenient to have (60) expressed in terms of (62), i.e.

$$\left.\begin{aligned}
\sigma_1^2 &= N[s^2(c)+m^2(c)]\,(1/p_1-p_2/p_1^2) + N^2 m^2(c)\,(p_2/p_1^2-1)\,, \\
\sigma_2^2 &= N[s^2(a)+m^2(a)]\,(p_3-p_4)/p_2^2+\binom{N}{2}^2 m^2(C)\,(p_4/p_2^2-1)+ \\
&\quad +\binom{N}{2}[s^2(C)+m^2(C)]\,(p_2-2p_3+p_4)/p_2^2\,, \\
\sigma_{12} &= N[s(a,c)+m(a)m(c)]\,(p_2-p_3)/p_1p_2+N\binom{N}{2}m(C)m(c)\,(p_3/p_1p_2-1).
\end{aligned}\right\} \quad (63)$$

The subgraph totals are symmetric functions of variables with double indices. In a paper by Zykov (1957) it is shown that the symmetric functions corresponding to the connected graphs act as an algebraic basis of all the polynomial symmetric functions of doubly indexed variables, i.e. the symmetric functions are uniquely decomposable in the same way as the graphs are uniquely decomposable into connected components.

The subgraph totals have been used by Barton and David (1966) in a study of random graphs proposed to analyze secular and geographic adjacencies of childhood leukemia. Bloemena (1964) has used the subgraph totals to find limit distributions of certain sample graph characteristics that are related to the estimation problems of the present section.

4.5 Unbiased estimators of the variances

The principal advantage of the subgraph totals is of a theoretical nature. It is very simple to get unbiased estimators of the subgraph totals of the population graph. Consider the corresponding subgraph totals of the sample graph. They are denoted by lower case letters and can be written in analogy with (61), as

$$\left.\begin{aligned}
t_{11} &= \sum_{\alpha} Z_{\alpha\alpha} = \sum_{i} C_{ii}^2\,\theta_i\,, \\
t_{12} &= 2\sum\sum_{\alpha<\beta} Z_{\alpha\alpha}\,Z_{\beta\beta} = 2\sum\sum_{i<j} C_{ii}C_{jj}\,\theta_i\theta_j\,, \\
t_{22} &= \sum\sum_{\alpha<\beta} Z_{\alpha\beta}^2 = \sum\sum_{i<j} C_{ij}^2\,\theta_i\theta_j\,, \\
&\text{etc.}
\end{aligned}\right\} \quad (64)$$

Hence it immediately follows that $E\, t_{uv} = T_{uv} p_v$, and an unbiased estimator of T_{uv} is obtained as $\hat{T}_{uv} = t_{uv}/p_v$. The variances and the covariance in (60) are linear functions of the T_{uv}, and consequently they have the unbiased estimators

$$\left.\begin{aligned} \hat{\sigma}_1^2 &= t_{11}(1/p_1^2 - 1/p_1) + t_{12}(1/p_1^2 - 1/p_2)\ , \\ \hat{\sigma}_2^2 &= t_{22}(1/p_2^2 - 1/p_2) + t_{23}(1/p_2^2 - 1/p_3) + t_{24}(1/p_2^2 - 1/p_4), \\ \hat{\sigma}_{12} &= t_{32}(1/p_1p_2 - 1/p_2) + t_{33}(1/p_1p_2 - 1/p_3)\ . \end{aligned}\right\} \qquad (65)$$

The subgraph totals t_{uv} of the sample graph can be expressed like the totals T_{uv} in (62). That is achieved by writing n, Z, x, and z instead of N, C, a, and c in the formulas (62). By substitution in (65) and simplification,

$$\left.\begin{aligned} \hat{\sigma}_1^2 &= n[s^2(z)+m^2(z)]\,(1/p_2 - 1/p_1) + n^2 m^2(z)\,(1/p_1^2 - 1/p_2), \\ \hat{\sigma}_2^2 &= n[s^2(x)+m^2(x)]\,(1/p_4 - 1/p_3) + \tbinom{n}{2} m^2(Z)\,(1/p_2^2 - 1/p_4) + \\ &\quad + \tbinom{n}{2}[s^2(Z)+m^2(Z)]\,(2/p_3 - 1/p_2 - 1/p_4)\ , \\ \hat{\sigma}_{12} &= n[s(x,z)+m(x)m(z)]\,(1/p_3 - 1/p_2) + n\tbinom{n}{2} m(Z)m(z)\,(1/p_1p_2 - 1/p_3). \end{aligned}\right\} \qquad (66)$$

These formulas provide unbiased estimators of the variances and the covariance of the estimators of the node and arc totals. In the next three subsections, some special cases will be studied. In Subsection 4.6 the relations to conventional sampling will be pointed out. In Subsection 4.7 the results will be particularized to the situation that was considered in Section 3. In Subsection 4.8 the estimation of $T = T_1 + T_2$ is considered as an illustration of the dependence between $\hat{T}_1$ and $\hat{T}_2$. Another illustration is provided by Subsection 6.6.

4.6 Estimation of the node total

If the node total T_1 is to be estimated from a node sample, the graph structure of the population is immaterial, and the problem is a conventional one.

The node total T_1 is estimated by $\hat{T}_1 = t_1/p_1$, which has the variance σ_1^2 given by (63). From (66) its unbiased estimator $\hat{\sigma}_1^2$ is obtained.

When the sample consists of n nodes chosen without replacement, it follows from (50) that $\hat{T}_1 = N\, m(z)$, and the following well-known formulas are applicable:

$$\sigma_1^2 = N^2(N-n)s^2(c)/(N-1)n\ , \tag{67}$$

$$\hat{\sigma}_1^2 = N(N-n)s^2(z)/(n-1)\ . \tag{68}$$

When Bernoulli sampling with selection probability $p = 1-q$ is used, it follows from (51) that $\hat{T}_1 = t_1/p$, and (63) simplifies to

$$\sigma_1^2 = N[s^2(c)+m^2(c)]\, q/p\ . \tag{69}$$

From (66) it follows that

$$\hat{\sigma}_1^2 = n[s^2(z)+m^2(z)]\, q/p^2. \tag{70}$$

Particularly if all the node values are equal to 1, an estimator $\hat{N} = n/p$ is obtained. Moreover, Var $\hat{N} = \sigma_1^2 = N\, q/p$ and $\hat{\sigma}_1^2 = n\, q/p^2$, which are well-known formulas for a stochastic variable n that is binomial (N, p).

4.7 Estimation of the arc total

The arc total T_2 is estimated by $\hat{T}_2 = t_2/p_2$, which has the variance σ_2^2 given by (63). From (66) its unbiased estimator $\hat{\sigma}_2^2$ is obtained.

If each arc value is equal to 0 or 1, the arc total T_2 becomes equal to the arc frequency R of a simple graph. Then the arc frequency R has the unbiased estimator $\hat{R} = r/p_2$, and the variance in (63) simplifies to

$$\sigma_2^2 = N[s^2(a)+4R^2/N^2]\,(p_3-p_4)/p_2^2+R^2(p_4/p_2^2-1)+R(p_2-2p_3+p_4)/p_2^2. \tag{71}$$

Its estimator simplifies to

$$\hat{\sigma}_2^2 = n[s^2(x)+4r^2/n^2](1/p_4-1/p_3)+r^2(1/p_2^2-1/p_4)+r(2/p_3-1/p_2-1/p_4). \quad (72)$$

If n nodes are sampled without replacement, it follows from (50) that R has the estimator $\hat{R} = N(N-1)r/n(n-1)$, and (71) and (72) become identical to (37) and (46) respectively. If Bernoulli sampling with selection probability $p=1-q$ is used it follows from (51) that R has the estimator $\hat{R} = r/p^2$, and (71) and (72) become simplified to

$$\sigma_2^2 = N[s^2(a)+4R^2/N^2]q/p+Rq^2/p^2 \quad (73)$$

and

$$\hat{\sigma}_2^2 = n[s^2(x)+4r^2/n^2]q/p^4-rq^2/p^4. \quad (74)$$

When $p>0$ and not all x_α are equal to 0, it follows that

$$\hat{\sigma}_2^2 = \sum_\alpha x_\alpha^2 q/p^4-\sum_\alpha x_\alpha q^2/2p^4 > \sum_\alpha x_\alpha(x_\alpha-1/2)q/p^4 > 0. \quad (75)$$

Thus, if Bernoulli sampling with $p = n/N$ is considered as an approximation to sampling of n nodes without replacement, it follows that $\hat{\sigma}_2^2 > 0$ when n is small compared with N.

4.8 Estimation of the sum of the node and arc totals

To demonstrate the use of the covariance estimator $\hat{\sigma}_{12}$, the sum $T = T_1+T_2$ of the node and arc totals will be estimated by $\hat{T} = \hat{T}_1+\hat{T}_2$. This estimator is unbiased, and its variance $\mathrm{Var}\,\hat{T} = \sigma^2 = \sigma_1^2+2\sigma_{12}+\sigma_2^2$ can be unbiasedly estimated by $\hat{\sigma}^2 = \hat{\sigma}_1^2+2\hat{\sigma}_{12}+\hat{\sigma}_2^2$.

If the node values are the circuit arc frequencies and the arc values are the arc frequencies, T becomes equal to the total arc frequency.

4.9 A priori knowledge about the population graph

The results so far are based upon the assumption that the population matrix C is symmetric but otherwise completely unknown. If there is available some a priori knowledge about the matrix C the use of it may give better estimators. To illustrate this, an extremely simple and special situation will be considered.

Suppose that the arc frequency R of an undirected graph without circuit arcs and without multiple arcs is to be estimated. A Bernoulli sample with selection probability p is available. The estimator $\hat{R} = r/p^2$ has a variance given by (73). Now suppose that it is known that the graph is a tree. Then it is true that R=N-1, and R can be estimated without bias by $R^* = \hat{N}-1$, where $\hat{N} = n/p$. The estimator R^* has the variance $\operatorname{Var} R^* = \operatorname{Var} \hat{N} = Nq/p$. A comparison with (73) shows that $\operatorname{Var} \hat{R} > \operatorname{Var} R^*$ when $N > 2$. Thus, the a priori knowledge implies that an estimator of smaller variance can be found.

4.10 A matrix sampling design

Subgraph sampling can be looked upon as a sampling of elements from matrices, i.e. sampling of doubly indexed units.

Let B be an arbitrary $N \times N$ matrix, and consider its elements B_{ij} as the units of a population. A sample is chosen from the set of indices $1,2,\ldots,N$, and the elements of B that correspond to the sampled rows and columns constitute a submatrix Y that is observed.

If the problem is to estimate the total $T = \sum_i \sum_j B_{ij}$, it is possible to introduce the symmetrized matrices $C = B+B'$ and $Z = Y+Y'$, and to write $T = T_1/2+T_2$ with $T_1 = \sum_i C_{ii}$ and $T_2 = \sum\sum_{i<j} C_{ij}$. The totals T_1 and T_2 can be estimated as in Subsection 4.3, and the variances and covariance of the estimators are given in Subsection 4.5.

There are instances where a sampling scheme of this kind is natural. Let B_{ij} be some property of a mixing of two objects i and j, for instance two drugs, two colors, etc. Then it may be advantageous for economical or experimental reasons to choose a sample of objects and make all paired mixings of the sampled objects.

4.11 A comparison between matrix sampling and ordinary sampling

Even in cases where the matrix sampling designs do not have a natural reason it may occasionally be possible to use this sampling method to get better estimators than with ordinary sampling. An example will be studied to demonstrate this fact.

Let a population consist of $\binom{N}{2}$ units whose values are denoted C_{ij} for $i = 1,\ldots,j-1$ and $j = 2,3,\ldots,N$. The values C_{ij} can be interpreted as the elements of a symmetric matrix C with zeros in the main diagonal. To estimate the total $T_2 = \sum\sum_{i<j} C_{ij}$, two sampling methods will be considered and compared, i.e. matrix sampling and ordinary sampling of units of the population. To simplify the formula writing, Bernoulli sampling schemes will be considered.

The selection probabilities will be chosen in such a way that the expected numbers of observed units will be the same in the two sampling schemes. This implies that if the matrix sampling is carried out as a Bernoulli sampling with the selection probability $p=1-q$, then the other sampling scheme has to be Bernoulli sampling with the selection probability p^2. In the two cases the expected number of observed units becomes $\binom{N}{2}p^2$.

When matrix sampling is used, the total T_2 is estimated by $\hat{T}_2 = t_2/p^2$, and the variance of the estimator is given according to (60) as

$$\text{Var } \hat{T}_2 = T_{22}(1-p^2)/p^2 + T_{23}\, q/p. \tag{76}$$

When sampling of units is used, the total T_2 is estimated by

$$T_2^* = \sum\sum_{i<j} C_{ij}\theta_{ij}/p^2, \tag{77}$$

where the sample selection indicator variables θ_{ij} are independent Bernoulli (p^2). Hence

$$\text{Var } T_2^* = \sum\sum_{i<j} C_{ij}^2(1-p^2)/p^2 = T_{22}(1-p^2)/p^2. \tag{78}$$

From (76) and (78) it follows that

$$\text{Var } \hat{T}_2 - \text{Var } T_2^* = T_{23}\, q/p, \tag{79}$$

with T_{23} defined as in (61) and (62). Consequently, $\hat{T}_2$ has a smaller variance than T_2^* if, and only if, the population parameter T_{23} is negative. A simple case where matrix sampling is preferable is, for instance, when C is checkerwise composed of the digits 1 and -1 outside the main diagonal. If N is odd, then $T_{23} = -(N-1)(N-2)$.

In practice the units of the population have to be ordered in the triangular fashion C_{ij} for $i<j$. If this ordering is looked upon as a randomization, it may be possible to state conditions for the matrix sampling design to be better on the average than the ordinary sampling. This topic will not be considered here.

5. ESTIMATION OF THE DISTRIBUTION OF THE LOCAL DEGREES

5.1 Introduction

Consideration has been given in Sections 3 and 4 above to the question of how to estimate the total arc frequency in directed or undirected graphs with or without circuit arcs and multiple arcs. When the node frequency N is known, it may be said that the results of Sections 3 and 4 refer to the mean values m(a), m(b), and m(c) of the number of out-arcs, in-arcs, and circuit arcs at the nodes.

A more far-reaching problem is to estimate the number of nodes that have fixed frequencies of out-arcs, in-arcs, and circuit arcs. This is a considerably more difficult problem, and in the following subsections some special cases will be studied. The achieved results are not so complete as the results of Section 3 or 4. Unbiased estimators will be given, and a method indicated for computation of the variances and the covariances of the estimators. However, these variances and covariances are not estimated by use of the sample graph.

5.2 The undirected case

For the simplest case of an undirected graph without circuit arcs and multiple arcs, the problem becomes one of estimating the number of nodes that have a fixed local degree, i.e. of estimating the frequencies $f_a(u)$ for $u = 0, 1, \ldots, N-1$.

The estimation is to be based upon a subgraph obtained by sampling n nodes without replacement from the N nodes of the population graph.

Intuitively, some information about the distribution of the local degrees in the population graph is provided by a study of the distribution of the local degrees in the sample graph. The observed frequencies $f_x(u)$ for $u = 0,1,\ldots,n-1$ will therefore by analyzed.

The expected value of the observed frequency $f_x(u)$ can be written

$$E\, f_x(u) = \sum_{\alpha=1}^{n} P(x_\alpha = u), \tag{80}$$

and the probability $P(x_\alpha = u)$ can be determined by a combinatorial argument. Node α of the sample graph equals node i of the population graph with the probability 1/N. Subject to choice α having resulted in node i, the remaining n-1 nodes of the sample graph can be chosen from the remaining N-1 nodes of the population graph in $\binom{N-1}{n-1}$ different ways. These n-1 nodes must consist of u nodes chosen from the a_i nodes that have an arc with node i, and n-1-u nodes chosen from the $N-1-a_i$ nodes that do not have an arc with node i. Consequently,

$$E\, f_x(u) = n \sum_{i=1}^{N} \binom{a_i}{u} \binom{N-1-a_i}{n-1-u} / N \binom{N-1}{n-1} =$$

$$= n \sum_{U=0}^{N-1} f_a(U) \binom{U}{u} \binom{N-1-U}{n-1-u} / N \binom{N-1}{n-1} \tag{81}$$

for $u = 0,1,\ldots,n-1$. It should be noticed that the sum that ranges over $u \le U \le N-n+u$ can be extended to range over $0 \le U \le N-1$ if the binomial coefficients are conventionally defined, i.e. $\binom{U}{u} = 0$ for $U < u$. Introduce the relative frequencies of the local degrees, and put for abbreviation

$$P(u,U) = \binom{U}{u} \binom{N-1-U}{n-1-u} / \binom{N-1}{n-1} \tag{82}$$

for $u = 0,1,\ldots,n-1$ and $U = 0,1,\ldots,N-1$. Then it follows from (81) that

$$E\, p_x(u) = \sum_{U=0}^{N-1} P(u,U)\, p_a(U) \tag{83}$$

for $u = 0,1,\ldots,n-1$. The equations (83) give the n expected frequencies $E\, p_x(u)$ as linear combinations of the N frequencies $p_a(U)$ of the population. When $n < N$, the equation system is under-determined. If a sufficient number of consistent conditions are imposed on the population frequencies, it may be possible to give the population frequencies $p_a(U)$ as linear combinations of the expected values $E\, p_x(u)$. If the sample frequencies $p_x(u)$ are substituted for these expected values, unbiased estimators $\hat{p}_a(U)$ of $p_a(U)$ will be obtained.

In many cases an appropriate approach to the problem is to put $p_a(U) = 0$ for $U \geq n$. This condition means that the nodes of the population graph are supposed to have at most as many arcs as are possible to observe in the sample graph. According to this assumption, the unbiased estimators $\hat{p}_a(U)$ are given by the equation system

$$p_x(u) = \sum_{U=0}^{n-1} P(u,U)\,\hat{p}_a(U) \tag{84}$$

for $u = 0,1,\ldots,n-1$. The equation system (84) is triangular and non-singular, i.e. $P(u,U) = 0$ for $U < u$, and all the diagonal elements $P(u,u)$ are positive. Then it is possible to successively solve for $\hat{p}_a(U)$ in the reversed order $U = n-1, n-2, \ldots, 0$. The solution equals

$$\hat{p}_a(U) = \sum_{u=0}^{n-1} Q(U,u)\,p_x(u) \tag{85}$$

for $U = 0,1,\ldots,n-1$, with

$$Q(U,u) = \binom{u}{U}\binom{n-1-u}{N-1-U} \Big/ \binom{n-1}{N-1} = \binom{N-1}{U}\binom{n-N}{u-U} \Big/ \binom{n-1}{u} \tag{86}$$

for $U = 0,1,\ldots,n-1$ and $u = 0,1,\ldots,n-1$. Here $Q(U,u) = 0$ for $u < U$, and $Q(U,U) > 0$ for $U = 0,1,\ldots,n-1$. That (85) is in fact a solution of (84) is seen from the following (compare Feller, 1957, Chapter 2):

$$\sum_U P(u,U)\,Q(U,v) = \sum_U \binom{U}{u}\binom{N-1-U}{n-1-u}\binom{N-1}{U}\binom{n-N}{v-U}\Big/\binom{N-1}{n-1}\binom{n-1}{v} =$$

$$= \sum_U \binom{N-n}{U-u}\binom{n-N}{v-U}\binom{n-1}{u}\Big/\binom{n-1}{v} = \binom{0}{v-u}\binom{n-1}{u}\Big/\binom{n-1}{v} = I(u=v). \qquad (87)$$

5.3 A comment about the covariances

To compute the variances and covariances of the estimators provided by (85), the variances and covariances of the sample frequencies $p_x(u)$ have to be considered. In fact, it follows from (85) that

$$\mathrm{Cov}\,[\hat{p}_a(U),\ \hat{p}_a(V)] = \sum_u \sum_v Q(U,u)\,Q(V,v)\ \mathrm{Cov}\,[p_x(u),\ p_x(v)]. \qquad (88)$$

Consider then

$$E\,p_x(u)\,p_x(v) = \sum_\alpha \sum_\beta P(x_\alpha = u,\ x_\beta = v)/n^2. \qquad (89)$$

By a combinatorial argument it is found that (89) equals

$$I(u=v)\,E\,p_x(u)/n +$$

$$+(n-1)\sum_{i\neq j}\sum_{w=0}^{N-2}\binom{B_{ij}}{w}\binom{a_i-C_{ij}-B_{ij}}{u-C_{ij}-w}\binom{a_j-C_{ij}-B_{ij}}{v-C_{ij}-w}\binom{N-2-a_i-a_j+2C_{ij}+B_{ij}}{n-2-u-v+2C_{ij}+w}\Big/$$

$$\Big/\binom{N-2}{n-2}nN(N-1), \qquad (90)$$

where

$$B_{ij} = \sum_{k=1}^{N} C_{ik}\,C_{jk}\,. \qquad (91)$$

This result if of no practical use, because it involves unknown population parameters that seem hard to estimate. To get any further, the simultaneous distribution of $(a_i, a_j, B_{ij}, C_{ij})$ has to be introduced and estimated from the sample graph. This procedure is undoubtedly an unattractive one, and it seems likely that it will be more fruitful to search for another approach to the problem.

5.4 The directed case

Consider a directed graph without circuit arcs and multiple arcs. The problem here becomes one of estimating $f_{a,b}(u,v)$ for $u = 0,1,\ldots,N-1$ and $v = 0,1,\ldots,N-1$. A more thorough characterization of the nodes consists not only of the out-degrees and the in-degrees but also the number of nodes that have arcs in the two directions. Let a pair of opposite directed arcs between two nodes be called a double arc, and let for the moment the frequency of double arcs at node i be denoted by c_i. The three characteristics a_i, b_i, and c_i of node i are sufficient to enable simultaneous distribution of the local out-arc and in-arc frequencies, i.e. the i-th row and column of the matrix C, to be obtained. Thus it is natural to consider the frequencies $f_{a,b,c}(u,v,w)$. The relation

$$f_{a,b}(u,v) = \sum_{w=0}^{N-1} f_{a,b,c}(u,v,w) \tag{92}$$

can be used to get the estimators of $f_{a,b}(u,v)$ from the estimators of $f_{a,b,c}(u,v,w)$.

To estimate $f_{a,b,c}(u,v,w)$, the corresponding frequencies $f_{x,y,z}(u,v,w)$ of the sample graph will be considered. Here z denotes the vector whose components are the double arc frequencies of the n nodes of the sample graph. It holds that

$$E\, f_{x,y,z}(u,v,w) = \sum_{\alpha=1}^{n} P(x_\alpha=u,\ y_\alpha=v,\ z_\alpha=w), \tag{93}$$

and according to a combinatorial argument of the kind preceding (81) it follows that

$$E\, f_{x,y,z}(u,v,w) = n \sum_{i=1}^{N} \binom{a_i-c_i}{u-w}\binom{b_i-c_i}{v-w}\binom{c_i}{w}\binom{N-1-a_i-b_i+c_i}{n-1-u-v+w} \Big/ \Big/ N\binom{N-1}{n-1}. \tag{94}$$

Introduce the relative frequencies and put

$$P(u,v,w,U,V,W) = \binom{U-W}{u-w}\binom{V-W}{v-w}\binom{W}{w}\binom{N-1-U-V+W}{n-1-u-v+w}\Big/\binom{N-1}{n-1}. \qquad (95)$$

Then (94) can be written as

$$E\, p_{x,y,z}(u,v,w) = \sum_U \sum_V \sum_W P(u,v,w,U,V,W)\, p_{a,b,c}(U,V,W). \qquad (96)$$

Suppose that $p_{a,b,c}(U,V,W) = 0$ if $U \geq n$ or if $V \geq n$. This means that the nodes of the population graph are assumed to have at most as many out-arcs, in-arcs, and double arcs that are possible to observe in the sample graph. This assumption is quite natural, as the sample size ought to be so large that all the frequencies that it is intended to estimate will have a chance of being represented in the sample graph.

According to this assumption, the unbiased estimators $\hat{p}_{a,b,c}(U,V,W)$ of $p_{a,b,c}(U,V,W)$ are obtained from the equation system

$$p_{x,y,z}(u,v,w) = \sum_{U=0}^{n-1} \sum_{V=0}^{n-1} \sum_{W=0}^{n-1} P(u,v,w,U,V,W)\, \hat{p}_{a,b,c}(U,V,W) \qquad (97)$$

for $0 \leq u-w$, $0 \leq v-w$, $0 \leq w$, $u+v-w \leq n-1$. The coefficients of the equation system (97) are zero if the following conditions are not simultaneously satisfied: $U-W \geq u-w$, $V-W \geq v-w$, $W \geq w$, $N-1-U-V+W \geq n-1-u-v+w$. As a consequence, it is possible to solve successively for $\hat{p}_{a,b,c}(U,V,W)$ in the reversed lexicographical order. The solution becomes

$$\hat{p}_{a,b,c}(U,V,W) = \sum_{u=0}^{n-1} \sum_{v=0}^{n-1} \sum_{w=0}^{n-1} Q(U,V,W,u,v,w)\, p_{x,y,z}(u,v,w) \qquad (98)$$

for $0 \leq U-W$, $0 \leq V-W$, $0 \leq W$, $U+V-W \leq n-1$, with the coefficients given by

$$Q(U,V,W,u,v,w) = \binom{u-w}{U-W}\binom{v-w}{V-W}\binom{w}{W}\binom{n-1-u-v+w}{N-1-U-V+W}\Big/\binom{n-1}{N-1}. \qquad (99)$$

The proof of (98) runs as the proof of (85), and the result is in complete analogy

with the result for the undirected graphs. The position in this case, even more than in the previous one, is that the computations of the variances and the covariances of the estimators are too complicated to be useful in practice.

5.5 Application to a problem in sociometry

A problem in sociology that has deserved some interest is to estimate the number of individuals that have no contacts with the other individuals of a certain group, i.e. the so called isolates of the group.

In papers by Edwards (1948) and Katz (1951) the approach to the problem was to use stochastic contacts to deduce the distribution of the number of isolates in the group. Such stochastic models will be taken up in Chapter 7. Another approach to the problem that does not seem to have been studied earlier is to use the information of a sampled subgroup to estimate the number of isolates in a population group. Obviously, in many practical cases this is what is really wanted, and the approach with stochastic contacts was merely a means of formulating the problem in a way that was appropriate for the mathematical treatment.

Let there be given an undirected graph without circuit arcs and multiple arcs. The node frequency N will be known. The proportion if isolates is $p_a(0)$, and according to (85) and (86) an unbiased estimator of the proportion $p_a(0)$ is provided by

$$\hat{p}_a(0) = \sum_{u=0}^{n-1} (-1)^u \binom{N-n+u-1}{u} p_x(u) / \binom{n-1}{u} , \qquad (100)$$

if it is assumed that the sample size n is larger than the maximum local degree of the population graph, i.e. $p_a(U) = 0$ for $U \geq n$. It is seen from (100) that the proportion of isolates in the population graph is estimated by an alternating series in the frequencies of the local degrees of the sample graph.

As a numerical illustration, consider the simple population graph of Subsection 2.5, i.e. the graph of Figure 1. The population graph has N=8, $f_a(0)=0$, and the maximum local degree is 4. If a sample of n=5 nodes is chosen,

$$\hat{p}_a(0) = p_x(0)-(3/4)p_x(1)+p_x(2)-(5/2)p_x(3)+15p_x(4). \quad (101)$$

From the graph list, Table 1, and from Table 2 it follows that the estimator $\hat{p}_a(0)$ takes the values reported in Table 14. The sampling distribution of the estimator $\hat{p}_a(0)$ is given in Table 15. Though the estimator is unbiased, i.e. $E\,\hat{p}_a(0)=0$, it behaves in a rather intractable manner. For instance it takes values outside the range from 0 to 1.

Table 14. The estimator $\hat{p}_a(0)$ corresponding to the sample size 5 and the population graph shown in Figure 1.

Sample graph	Frequency	x	Distribution of x	$\hat{p}_a(0)$
5,4,c	8	02222	10400	1
5,4,f	8	11222	02300	0.3
5,5,c	8	11233	02120	-1.1
5,4,e	6	11123	03110	-0.75
5,2,b	4	00112	22100	0.3
5,3,b	4	01113	13010	-0.75
5,3,c	4	11112	04100	-0.4
5,5,b	4	12223	01310	-0.05
5,5,a	2	11224	02201	3.1
5,6,a	2	22233	00320	-0.4
5,6,c	2	12234	01211	2.75
5,6,e	2	22233	00320	-0.4
5,7,b	2	23333	00140	-1.8
	56			

Table 15. The sampling distribution of the estimator $\hat{p}_a(0)$ corresponding to the sample size 5 and the population graph shown in Figure 1.

$\hat{p}_a(0)$	Frequency
-1.8	2
-1.1	8
-0.75	10
-0.4	8
-0.05	4
0.3	12
1	8
2.75	2
3.1	2
	56

6. ESTIMATION OF THE NUMBER OF CONNECTED COMPONENTS

6.1 Notations

Consider an undirected population graph without circuit arcs and multiple arcs. The graph consists of K connected components that have $N_1, N_2, \ldots, N_K$ nodes. The graph has $N=N_1+N_2+\ldots+N_K$ nodes in total. Let H_U denote the number of components that consist of exactly U nodes for $U=1,2,\ldots,N$. Consequently,

$$K = \sum_{U=1}^{N} H_U \quad \text{and} \quad N = \sum_{U=1}^{N} U H_U . \tag{102}$$

Let U_i denote the node frequency of the connected component containing node i.

The number U_i is called the reach of node i. As all the nodes in the same connected component with U nodes have the same value U of their reaches, and as there are H_U connected components of U nodes, it follows that UH_U nodes have the reach U for U=1,2,...,N. Then, according to (102), the number K of connected components can be written as

$$K = \sum_{i=1}^{N} 1/U_i \,. \tag{103}$$

Thus, the mean number of nodes per connected component, i.e. N/K, can be interpreted as the harmonic mean value of the reaches of the nodes. As the harmonic mean is less than or equal to the arithmetic mean with equality if, and only if, all the terms are equal,

$$K \geq N^2 / \sum_{i=1}^{N} U_i = N^2 / \sum_{U=1}^{N} U^2 H_U = N^2 / \sum_{v=1}^{K} N_v^2 \,, \tag{104}$$

with equality if, and only if, all the connected components consist of the same number of nodes.

From the N nodes a random sample of n nodes is chosen without replacement. The subgraph corresponding to the sample is observed. For each node of the sample the reach within the sample can be observed, but not the reach in the population. Thus, two nodes may be connected in the population but unconnected in the sample. Let the number of connected components of the sample graph be denoted by k, and let the node frequencies of the connected components be equal to $n_1, n_2, \ldots, n_k$. Then $n=n_1+n_2+\ldots+n_k$. The number of connected components with u nodes in the sample graph is denoted by h_u for u=1,2,...,n. As for the population graph, it will be true that

$$k = \sum_{u=1}^{n} h_u \quad \text{and} \quad n = \sum_{u=1}^{n} u\, h_u \,. \tag{105}$$

Every connected component with exactly u nodes in the sample graph is a sub-

graph of a connected component with at least u nodes in the population graph. Several connected components of the sample graph may belong to the same connected component in the population graph. Consequently, k may be greater than K. All the connected components of the population graph need not be represented in the sample graph. Consequently, k may also be less than K.

6.2 Complete components

If all the connected components of the population graph are complete, i.e. have arcs between all their node pairs, the situation becomes considerably more simple. Then each connected component of the population graph can give rise to at most one connected component of the sample graph. In such a case k is not greater than K.

In this case, k can be interpreted by an urn model. An urn contains N balls of which N_1 have the color 1, N_2 have the color 2, ..., N_K have the color K. In a random sample of balls there are k different colors represented. Suppose that x_1 balls of the sample have the color 1, x_2 have the color 2, ..., x_K have the color K. All the positive x_v are observed and it is unknown how many of the x_v equal zero. The number of positive x_v is equal to k. If the same ball cannot enter the sample more than once, the positive numbers among $x_1, x_2, \ldots, x_K$ are equal to the numbers $n_1, n_2, \ldots, n_k$, i.e. the node sizes of the connected components of the sample graph.

The probability distribution of the stochastic variables $x_1, x_2, \ldots, x_K$ is easily obtained for some ordinary sampling schemes. If n balls are selected by sampling without replacement, the outcome $x_1, x_2, \ldots, x_K$ has the probability

$$\prod_{v=1}^{K} \binom{N_v}{x_v} \Big/ \binom{N}{n} . \tag{106}$$

If Bernoulli sampling with selection probability p=1-q is used, the outcome $x_1, x_2, \ldots, x_K$ has the probability

$$\prod_{v=1}^{K} \binom{N_v}{x_v} p^{x_v} q^{N_v - x_v} . \tag{107}$$

If m balls are selected by sampling with replacement, the outcome $x_1, x_2, \ldots, x_K$ has the probability

$$(m! / \prod_{v=1}^{K} x_v!) \prod_{v=1}^{K} (N_v/N)^{x_v} . \tag{108}$$

As $k \leq K$ when the connected components are complete, it follows that $E\,k \leq K$. If n nodes are sampled without replacement,

$$E\,k = \sum_{v=1}^{K} P(x_v > 0) = \sum_{v=1}^{K} \left[1 - \binom{N-N_v}{n} / \binom{N}{n} \right] , \tag{109}$$

and hence $E\,k = K$ if, and only if, n is larger than $N - \min N_v$. Especially if there is any isolated node in the population graph, the condition of unbiasedness implies that the sample has to comprise all the nodes of the population.

If the sampling follows a Bernoulli scheme with selection probability $p = 1-q$, the expected value is

$$E\,k = \sum_{v=1}^{K} (1-q^{N_v}) , \tag{110}$$

which is less than K if $p < 1$. Thus, k has negative bias as an estimator of K when $p < 1$.

If m nodes are sampled with replacement, then

$$E\,k = \sum_{v=1}^{K} \left[1 - (1-N_v/N)^m \right] , \tag{111}$$

which is less than K if $\min N_v < N$. Thus, k has negative bias as an estimator of K when there are at least two connected components in the population graph.

Generally, k provides a biased estimator of K. To get an unbiased estimator of K, it is not only k that has to be used, but also the values $n_1, n_2, \ldots, n_k$. Unbiased estimation of K will be considered for sampling without replacement

in the next subsection, after which Bernoulli sampling will be considered in Subsection 6.4.

6.3 Sampling without replacement

When the sample consists of n nodes chosen without replacement, then the number of connected components with u nodes in the sample graph has the following expected value

$$E\, h_u = \sum_{v=1}^{K} P(u, N_v) = \sum_{U=1}^{N} P(u, U)\, H_U , \tag{112}$$

with

$$P(u, U) = \binom{U}{u} \binom{N-U}{n-u} / \binom{N}{n} \tag{113}$$

for $u=1,2,\ldots,n$ and $U=1,2,\ldots,N$. The equation system (112) is of the same kind as the system (83) in Section 5. To get unbiased estimators $\hat{H}_U$ of H_U, the arguments of Section 5 may be repeated. Suppose that $H_U=0$ for $U > n$, i.e. suppose that the sample size n is at least equal to the node frequency of the largest connected component of the population graph. This implies that each connected component of the population graph will have a real chance of appearing as a connected component in the sample graph. With this assumption, the following equation system is obtained to determine the estimators $\hat{H}_U$ for $U = 1,2,\ldots,n$:

$$h_u = \sum_{U=1}^{n} P(u, U)\, \hat{H}_U \tag{114}$$

for $u=1,2,\ldots,n$. The equation system (114) is triangular and non-singular and can be successively solved for $\hat{H}_U$ in the reversed order $U=n,n-1,\ldots,1$. The solutions can be written

$$\hat{H}_U = \sum_{u=1}^{n} Q(U,u)\, h_u \,, \tag{115}$$

with

$$Q(U,u) = \binom{u}{U} \binom{n-u}{N-U} / \binom{n}{N} = \binom{N}{U} \binom{n-N}{u-U} / \binom{n}{u} \tag{116}$$

for $U=1,2,\ldots,n$ and $u=1,2,\ldots,n$. According to (102) and (115) it is found that K has the unbiased estimator

$$\hat{K} = \sum_{U=1}^{n} \hat{H}_U = \sum_{u=1}^{n} q_u\, h_u \,, \tag{117}$$

with $q_u = 1-Q(0,u)$, and $Q(0,u)$ defined by (116). N has the unbiased estimator

$$\hat{N} = \sum_{U=1}^{n} U\, \hat{H}_U = N \,. \tag{118}$$

Thus, the estimator $\hat{N}$ is equal to the known constant N, i.e. the estimators $\hat{H}_U$ are consistent with the node condition (102).

The variance of the estimator $\hat{K}$ can be derived from the variances and covariances of the variables h_u. According to (117),

$$\operatorname{Var} \hat{K} = \sum_u \sum_v q_u\, q_v \operatorname{Cov}(h_u, h_v) \,, \tag{119}$$

and

$$\operatorname{Cov}(h_u, h_v) = \sum_s \sum_t P(x_s = u,\, x_t = v) - E\, h_u\, E\, h_v =$$

$$= I(u=v)\, E\, h_u + \sum_{s \neq t} \sum P(x_s = u,\, x_t = v) - E\, h_u\, E\, h_v \,, \tag{120}$$

where

$$\sum_{s\neq t}\sum P(x_s=u,\ x_t=v) = \sum_{s\neq t}\sum \binom{N_s}{u}\binom{N_t}{v}\binom{N-N_s-N_t}{n-u-v}\Big/\binom{N}{n} =$$

$$= \sum_U \sum_V H_U H_V \binom{U}{u}\binom{V}{v}\binom{N-U-V}{n-u-v}\Big/\binom{N}{n} - \sum_U H_U \binom{U}{u}\binom{U}{v}\binom{N-2U}{n-u-v}\Big/\binom{N}{n}. \tag{121}$$

According to (119) - (121), the variance of $\hat{K}$ is given as a function of the population frequencies $H_1, H_2, \ldots, H_n$. In principle, an estimator of Var $\hat{K}$ is obtained by substituting $\hat{H}_1, \hat{H}_2, \ldots, \hat{H}_n$ for $H_1, H_2, \ldots, H_n$. Instead of going into these cumbersam calculations and studying the properties of the proposed estimator of Var $\hat{K}$, it is more tractable to study Bernoulli sampling.

6.4 Bernoulli sampling

If a Bernoulli sampling with selection probability p=1-q is used, then (112) holds with

$$P(u,U) = \binom{U}{u} p^u q^{U-u} \tag{122}$$

for u=1,2,...,N and U=1,2,...,N. The unbiased estimators $\hat{H}_U$ of H_U are given by the equation system

$$h_u = \sum_{U=1}^{N} P(u,U)\, \hat{H}_U \tag{123}$$

for u=1,2,...,N, which has the solution

$$\hat{H}_U = \sum_{u=1}^{N} Q(U,u)\, h_u\ , \tag{124}$$

with

$$Q(U,u) = \binom{u}{U} (1/p)^U (1-1/p)^{u-U} \tag{125}$$

for U=1,2,...,N and u=1,2,...,N. Then it is found that K has the estimator

$$\hat{K} = \sum_{U=1}^{N} \hat{H}_U = \sum_{u=1}^{N} q_u h_u , \tag{126}$$

with $q_u = 1 - Q(0,u)$, and $Q(0,u)$ defined by (125). N has the estimator

$$\hat{N} = \sum_{U=1}^{N} U \hat{H}_U = n/p . \tag{127}$$

The estimator $\hat{N}$ is the standard one based upon the sample size n that is binomial (N,p).

To calculate tha variance of $\hat{K}$ it is possible to proceed in exactly the same way as in the previous subsection, i.e. to use (119) and (120), where

$$\sum_{s \neq t} \sum P(x_s = u, x_t = v) = \sum_{s \neq t} \sum P(u, N_s) P(v, N_t) =$$

$$= E\, h_u \; E\, h_v - \sum_{U=1}^{N} P(u,U) P(v,U) H_U . \tag{128}$$

It follows that

$$\mathrm{Cov}\,(h_u, h_v) = \sum_{U=1}^{N} P(u,U) \left[I(u=v) - P(v,U) \right] H_U . \tag{129}$$

By using that

$$\sum_{u=1}^{N} q_u P(u,U) = 1 \text{ and } \sum_{u=1}^{N} q_u^2 P(u,U) = 1 + (q/p)^U , \tag{130}$$

it follows according to (119) and (129) that

$$\sigma^2 = \mathrm{Var}\, \hat{K} = \sum_{u=1}^{N} q_u (q_u - 1)\, E\, h_u = \sum_{U=1}^{N} (q/p)^U H_U \tag{131}$$

... u=1,2....,N. Thus, an unbiased estimator $\hat{\sigma}^2$ of σ^2 is given by

$$\hat{\sigma}^2 = \sum_{u=1}^{N} q_u(q_u-1)\, h_u = \sum_{u=1}^{N} (-q/p)^u \left[(-q/p)^u - 1\right] h_u \,. \tag{132}$$

There is an alternative way of obtaining these results which will also be described, since it has independent interest by giving $\hat{K}$ as a sum of independent stochastic variables. Write

$$h_u = \sum_{U=u}^{N} Y_{uU} \,, \tag{133}$$

with

$$Y_{uU} = \sum_{s=1}^{K} I(x_s=u,\, N_s=U) \tag{134}$$

for $u=1,2,\ldots,N$ and $U=1,2,\ldots,N$. The variable Y_{uU} denotes the number of connected components with u nodes in the sample graph which are subgraphs of any of the connected components with U nodes in the population graph. Then, according to (126) and (133)

$$\hat{K} = \sum_{U=1}^{N} y_U \,, \tag{135}$$

with

$$y_U = \sum_{u=1}^{U} q_u\, Y_{uU} \tag{136}$$

for $U=1,2,\ldots,N$.

As x_s are independent for different s values, and as $P(x_s=u) = P(u,U)$ when $N_s=U$, it follows that the vector variable $(Y_{0U}, Y_{1U}, \ldots, Y_{UU})$ is multinomial $(H_U,\ P(0,U),\ P(1,U),\ldots,\ P(U,U))$. Here $P(u,U)$ is given by (122) for $u=0,1,\ldots,U$ and $U=1,2,\ldots,N$. Moreover, the vector variables that belong to different U values are independent. Hence it follows that the variables y_U given by (136) are independent stochastic variables. According to well-known proper-

ties of the multinomial distribution (Cramér, 1947) it follows that

$$E\, y_U = \sum_u q_u\, P(u,U)\, H_U = H_U\,, \tag{137}$$

and

$$\mathrm{Var}\, y_U = \sum_u \sum_v q_u\, q_v \left[I(u=v)\, P(u,U) - P(u,U)\, P(v,U) \right] H_U = (q/p)^U H_U. \tag{138}$$

Hence it follows according to (135) that $\hat{K}$ is as unbiased estimator of K with the variance given by (131).

6.5 An example

To illustrate the use of the formulas above, a simple example will be considered.

The population graph is undirected. The connected components are complete and consist of 1, 2, or 3 nodes each. There are H_1, H_2, and H_3 connected components of the three kinds, respectively. There are $K = H_1+H_2+H_3$ connected components, $N = H_1+2H_2+3H_3$ nodes, and $R = H_2+3H_3$ arcs. A graph of this type is shown in Figure 3.

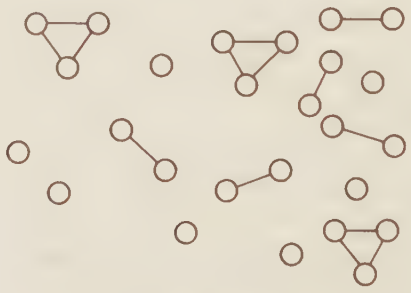


Figure 3

In a Bernoulli sample with selection probability p=1-q there are h_1, h_2, and h_3 connected components with 1, 2, and 3 nodes respectively. As a numerical illustration, let p = 0.25. According to (125) the numbers Q(U,u) are given by

the elements of the matrix

$$Q = \begin{vmatrix} 1/p & -2q/p^2 & 3q^2/p^3 \\ 0 & 1/p^2 & -3q/p^3 \\ 0 & 0 & 1/p^3 \end{vmatrix} = \begin{vmatrix} 4 & -24 & 108 \\ 0 & 16 & -144 \\ 0 & 0 & 64 \end{vmatrix} , \tag{139}$$

and the column sums are $q_1=4$, $q_2=-8$, and $q_3=28$. According to (124) it follows that

$$\left.\begin{aligned} \hat{H}_1 &= 4h_1 - 24h_2 + 108h_3 , \\ \hat{H}_2 &= 16h_2 - 144h_3 , \\ \hat{H}_3 &= 64h_3 , \end{aligned}\right\} \tag{140}$$

and according to (126) and (132)

$$\hat{K} = 4h_1 - 8h_2 + 28h_3 , \tag{141}$$

$$\hat{\sigma}^2 = 12h_1 + 72h_2 + 756h_3 . \tag{142}$$

6.6 Tree components

All the results of the present section obtained so far are based upon the assumption that there are only complete connected components. Without that assumption, the situation may be considerably more complicated. However, it is possible for certain special classes of graphs with incomplete connected components to get results of a character similar to those derived above.

As an example, we will consider the class of forest population graphs. The population graph is then undirected and has connected components that are tree graphs, i.e. without circuit paths. Let the population graph consist of H_1 isolated nodes, H_2 trees with two nodes each, H_3 trees with three nodes each, etc. The total number of nodes becomes $N = H_1+2H_2+3H_3+\ldots$, and the total number of arcs becomes $R = H_2+2H_3+3H_4+\ldots$. The number of connected com-

ponents is $K = H_1+H_2+H_3+\ldots = N-R$. A graph of this kind is shown in Figure 4.

From the results of Section 4 it is clear that an unbiased estimator of K is $\hat{K} = \hat{T}_1-\hat{T}_2$, where $\hat{T}_1=\hat{N}=n/p_1$ and $\hat{T}_2=\hat{R}=r/p_2$. The variance of K is obtained as $\sigma^2 = \text{Var } \hat{K}=\sigma_1^2+\sigma_2^2-2\sigma_{12}$ and estimated by $\hat{\sigma}^2=\hat{\sigma}_1^2+\hat{\sigma}_2^2-2\hat{\sigma}_{12}$. Here $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$,

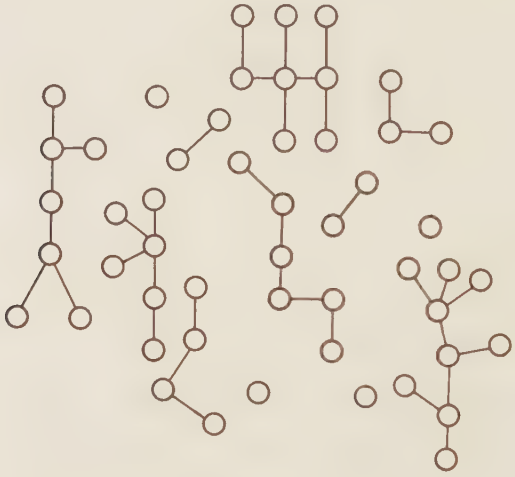


Figure 4

Figure 5

and $\hat{\sigma}_{12}$ are given by (66) after the appropriate simplifications that are due to the fact that all the node values are equal to 1, and all the arc values are equal to 0 or 1.

As a further specialization, consider the tree graphs that are straight chains of arcs. An example is shown in Figure 5. Each one of the H_U connected components with U nodes has (for $U > 1$) two nodes of local degree 1 and U-2 nodes of local degree 2. Then it follows by simple calculations that the formulas (66) can be given as functions of n, r, and h_1 only. Thus, the same property belongs to $\hat{\sigma}^2$. The details are omitted.

6.7 Comments

The results concerning the estimation of the number of connected components of a graph that are given above refer to complete components or to tree components only. In most applications, these cases are too special. The treatment may serve as an introduction to the general problem of estimating component frequencies by the use of a sampled subgraph. Maybe the account will stimulate to further work in this fascinating field of research.

If the methodology can be developed to include more general graphs, it is likely to have great potential for application in bacteriology, in particle physics, and in other areas where clumping of objects is studied. A book by Roach (1968) which deals with clumping of objects uses quite different stochastic models and does not enter into the sampling problems. However, the book has interest also for the sampling problems through its examples from diverse fields of application.

Chapter 5

INFERENCE FROM SAMPLED PARTIAL GRAPHS

SYNOPSIS

1. Sampled partial graphs. Some different kinds of partial graph are introduced. A useful connection between partial graphs and subgraphs is described. The sampling distribution of partial graphs is illustrated by an example.

2. Estimation of the arc frequency. The arc frequency is estimated in two ways for an undirected simple population graph. These two ways have complete counterparts in the directed case if one observes either only the out-arcs or both the out-arcs and the in-arcs at the nodes in the sample. Comparisons between subgraph and partial graph inference are commented upon.

3. Estimation of the node and arc totals. The general setup from Chapter 4, Section 4 is used for partial graph inference. Formulas are given for the variances and the covariance of the estimators. Unbiased variance and covariance estimators are provided.

4. Estimation of the distribution of the local degrees. For certain kinds of partial graph the estimation of the distribution of the local degrees is readily identified as a conventional problem. An interesting case requiring new methods

is that where the population graph is directed, the out-arcs from the nodes of the sample are observed, and the distribution of the local in-degrees is to be estimated. This case is discussed to some extent and is illustrated by a numerical example and a short account of various fields of application.

5. Estimation of the number of connected components. The estimation problem is shown to be capable of solution in a simple way when the connected components of the population graph are complete.

1. SAMPLED PARTIAL GRAPHS

1.1 A partial graph obtained by node sampling

Consider an undirected graph with N nodes as a population graph. A sample of n nodes is selected from the node set of this graph. All the arcs of each of the nodes in the sample are observed, and in this way the partial graph that is associated with the sampled node set is observed. It is important to note that the nodes of the partial graph which make up the sample are known. This knowledge makes it possible to separate the arcs between two nodes in the sample from the arcs between a node in the sample and a node outside the sample. If two nodes in the sample have no arc between them in the sample graph, they will not, as a consequence, have an arc between them in the population graph either.

In principle it might be imagined that the partial graph is observed without any knowledge of the nodes which constitute the sample. This case will not, however, be dealt with in the present chapter.

If C is the arc frequency matrix of the population graph, it is possible to describe the sampling and observation procedure in the following way. n out of the N rows of C are selected by random sampling. All the elements in the sampled rows are observed. As C is a symmetric matrix, consideration of columns instead of rows will yield equivalent results.

There exists a connection between partial graphs and subgraphs. The observed partial graph consists of all the arcs of the population graph that do not belong to the subgraph associated with the complement of the node sample. Consequently, the complement of the arc set of the partial graph is equal to the arc set of the subgraph associated with the complement of the node sample. This relation between partial graphs and subgraphs may occasionally be useful in getting results for sampled partial graphs from results for sampled subgraphs. Examples of this will be demonstrated below.

1.2 Generalization to directed graphs

When the population graph is directed it is possible to distinguish between two kinds of partial graph obtained by node sampling and application of different observation procedures. The first kind of partial graph is obtained when all the out-arcs from the sampled nodes are observed. The second kind of partial graph is obtained when both the out-arcs and the in-arcs associated with the sampled nodes are observed.

A partial graph that corresponds to the observation of all the in-arcs to the sample nodes may be considered to be a partial graph of the first kind, if the directions of all the arcs are changed. Consequently, it is unnecessary to deal with the case where only in-arcs are observed.

The two kinds of partial graph can be described in the following way by use of the arc frequency matrix C. A sample of n rows is selected out of the N rows of C, and in the first case all the elements that belong to the sampled rows are observed. In the second case a sample of n elements is selected out of the N elements in the main diagonal of C, and all the elements that belong to the same row or column as any of the sampled diagonal elements are observed.

For directed graphs there exists a connection between subgraphs and partial graphs of the second kind. A partial graph of the second kind has an arc set that is equal to the complement of the arc set of the subgraph associated with the complement of the sampled node set.

1.3 An example of a sampling distribution of partial graphs

An example will be given that, in spite of its simplicity, may be sufficient to familiarize the sampling distributions of partial graphs.

Consider the graph labeled 5,6,d in the graph list (Table 1) in Chapter 4, Subsection 2.3 as a population graph. Let the sample size be $n=2$. There are then $\binom{N}{n} = 10$ different samples. If the nodes of the population graph are labeled according to Figure 1, the partial graphs that correspond to the different samples can be given as in Table 1. The partial graphs are denoted by the labels of the above graph list. The sampling distribution of the partial graphs is readily obtained from Table 1 and is given in Table 2.

Table 1. Partial graphs corresponding to the samples with two nodes chosen from the population graph of Figure 1.

Sample	Partial graph
12	5,5,a
13	5,5,a
14	5,5,a
15	5,5,a
23	5,3,a
24	5,4,f
25	5,4,f
34	5,4,f
35	5,4,f
45	5,3,a

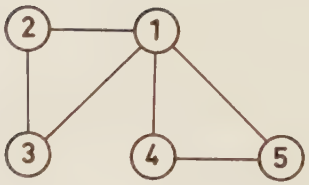


Figure 1

Table 2. Sampling distribution of the partial graphs obtained by sampling two nodes from the population graph of Figure 1.

Partial graph	Frequency
5,3,a	2
5,4,f	4
5,5,a	4

2. ESTIMATION OF THE ARC FREQUENCY

2.1 The undirected case

Let the population graph be undirected and consist of N nodes and R arcs. No circuit arcs and no multiple arcs are present. The notations introduced in Chapter 4, Subsection 1.2 will be used.

A random sample of n nodes is selected without replacement from the set of nodes. The associated partial graph is observed. The sample graph has N nodes and r arcs. By use of the sample selection indicators of Chapter 4, Subsection 2.7, the arc frequency r can be written

$$r = \sum_{i<j}\sum C_{ij}(\theta_i + \theta_j - \theta_i\theta_j). \qquad (1)$$

By simple transformations of (1), or by use of the connection between subgraphs and partial graphs, it follows that

$$r = R - \sum_{i<j} \sum C_{ij}(1-\theta_i)(1-\theta_j). \tag{2}$$

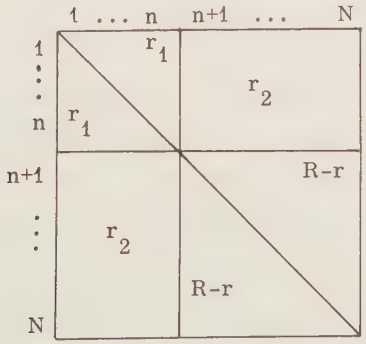
By substituting expected values in (1) or (2), it easily follows that R has the unbiased estimator

$$\hat{R} = r\, N(N-1)/n(2N-n-1). \tag{3}$$

To obtain the variance of $\hat{R}$, calculations of the same sort as in Chapter 4 may be used. From (2) it follows, however, that it is also possible to use (36) in Chapter 4 with N-n substituted for n. In this way, the following result appears after simplification

$$\begin{aligned}\operatorname{Var} \hat{R} = {} & N^2(N-1)(N-n)(N-n-1)(N-n-2)s^2(a)/(N-2)(N-3)(2N-n-1)^2 n + \\ & + N^2(N-1)^2(N-n)(N-n-1)(n-1)s^2(C)/2(N-2)(N-3)(2N-n-1)^2 n.\end{aligned} \tag{4}$$

Another estimator, R^*, of R can be found in the following way. The arc frequency r of the partial graph is written as the sum $r = r_1 + r_2$, where r_1 is the number of arcs between nodes in the sample, and r_2 is the number of arcs between the sample and its complement. If the nodes are labeled in such a way that the sample comprises the nodes $1, 2, \ldots, n$, the various arc frequencies can be represented as sums in the matrix C according to



Consider the variable

$$2r_1+r_2 = \sum_{i=1}^{N}\sum_{j=1}^{N} C_{ij}\theta_i = \sum_{i=1}^{N} a_i\theta_i , \tag{5}$$

which may be interpreted as the sum of n numbers $x_1, x_2, \ldots, x_n$ chosen without replacement from the N components $a_1, a_2, \ldots, a_N$ of the vector a. Let x be the vector of these numbers $x_1, x_2, \ldots, x_n$, and put $2r_1+r_2 = n\, m(x)$. From the well-known formulas $E\, m(x) = m(a) = 2R/N$ and $Var\, m(x) = (N-n)s^2(a)/(N-1)n$, it then follows that R has the unbiased estimator $R^* = N\, m(x)/2$, with the variance equal to

$$\sigma^2 = Var\, R^* = N^2(N-n)s^2(a)/4(N-1)n. \tag{6}$$

An unbiased estimator $\hat{\sigma}^2$ of σ^2 is obtained in the ordinary manner. For sampling without replacement, it holds that

$$E\, n\, s^2(x)/(n-1) = Ns^2(a)/(N-1), \tag{7}$$

and it follows that

$$\hat{\sigma}^2 = N(N-n)s^2(x)/4(n-1). \tag{8}$$

A comparison between the two variances given by (4) and (6) shows that

$$Var\, \hat{R}/Var\, R^* = 4(N-1)^2(N-n-1)(N-n-2)/(N-2)(N-3)(2N-n-1)^2 +$$
$$+ 2(N-1)^3(N-n-1)(n-1)s^2(C)/(N-2)(N-3)(2N-n-1)^2 s^2(a) . \tag{9}$$

A lower bound of the variance quotient is obtained by use of the inequality (15) of Chapter 4. If n and N tend to infinity in such a way that the fraction n/N tends to a fixed proportion p, it follows that the lower bound of the variance quotient tends to $(1-p)^2/(1-p/2)^2$. In particular, the position for $p = 0$ is that the variance quotient is greater than a number that tends to 1 for increasing N. As a consequence, it may be said that when n is small compared to N, the estimator R^* is generally preferred to $\hat{R}$ or equivalent to $\hat{R}$.

It may be added that the two estimators R^* and $\hat{R}$ are both linear combinations of r_1 and r_2. If one tries to determine the linear combination of r_1 and r_2 that has the least variance among all the unbiased linear combinations, one will find that the optimum coefficients depend upon certain population parameters. In order to use the optimum estimator, one either needs a priori knowledge about the values of these parameters or has to apply some two-stage estimation procedure.

The two estimators R^* and $\hat{R}$ which have been studied here have their counterparts in the directed case. R^* corresponds to the observation of out-arcs and $\hat{R}$ corresponds to the observation of both out-arcs and in-arcs.

2.2 Observation of out-arcs

Suppose that the population graph is directed and consists of N nodes and R arcs. No circuit arcs and multiple arcs are present. The notations of Chapter 4, Subsection 1.2 will be used.

A random sample of n nodes is chosen without replacement from the node set, and all the out-arcs from the sampled nodes are observed. The sample graph has N nodes and r arcs. By using the sample selection indicators, the arc frequency can be written

$$r = \sum_{i=1}^{N} \sum_{j=1}^{N} C_{ij} \theta_i = \sum_{i=1}^{N} a_i \theta_i , \tag{10}$$

and it can be considered as the sum of n numbers $x_1, x_2, \ldots, x_n$ chosen without replacement from the components of the vector a. Let x be the vector of these numbers, and put $r = n\, m(x)$. By a standard reasoning analogous to the one in the previous subsection it follows that R has the unbiased estimator $R^* = N\, m(x) = Nr/n$, with the variance

$$\sigma^2 = \mathrm{Var}\, R^* = N^2(N-n)s^2(a)/(N-1)n . \tag{11}$$

An unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = N(N-n)s^2(x)/(n-1). \tag{12}$$

2.3 Observation of out-arcs and in-arcs

If the assumptions of the preceding subsection are modified in such a way that both out-arcs and in-arcs are observed at the n nodes of the sample, the arc frequency of the sample graph can be given as

$$r = \sum_{i=1}^{N}\sum_{j=1}^{N} C_{ij}(\theta_i+\theta_j-\theta_i\theta_j) = R - \sum_{i=1}^{N}\sum_{j=1}^{N} C_{ij}(1-\theta_i)(1-\theta_j). \tag{13}$$

By taking expectations it readily follows, just as in Subsection 2.1, that R has the unbiased estimator

$$\hat{R} = r\, N(N-1)/n(2N-n-1). \tag{14}$$

However, the variance of $\hat{R}$ cannot be obtained by the argument that was used in Subsection 2.1. Due to the fact that C is not symmetric, the calculations needed to obtain the variance become more cumbersome. Instead of entering such calculations, the result may be obtained as a special case of a more general result of Section 3. To achieve this, the arc frequency is written

$$r = \sum_{i<j}\sum (C_{ij}+C_{ji})(\theta_i+\theta_j-\theta_i\theta_j). \tag{15}$$

This means that the population graph is considered to be undirected and without circuit arcs, but with multiple (double) arcs allowed. The treatment of this case is included in the more general problem that is studied in Section 3.

2.4 Comparisons between subgraph and partial graph inference

In cases where it is possible to choose between subgraph and partial graph sampling, a choice may be justified for technical reasons related to the investigation.

Cases may be imagined, for instance, where the nodes are individuals and the arcs are contacts of some specified kind. If the information about the contacts of the sampled individuals is gathered by interviews, the subgraph ob-

servation means that each sampled individual mentions his contacts with the other individuals of the sample only. If he also mentions his contacts with the individuals outside the sample, the information provides a partial graph. If there are many contacts, it may be difficult for the individuals who are interviewed to remember and to report all their contacts. For reliability reasons it may be appropriate to make each interviewed person mark his contacts on a list comprising only the individuals of the sample.

If the information about the contacts of the sampled individuals is recorded by some registration device that actually observes the contacts, it may be natural to use all the observed contacts of the sampled individuals and not only the contacts connecting two sampled individuals.

It may in some cases be appropriate to compare the efficiency of different sampling and observation procedures. Consider a case where the total cost of sampling investigation of a population graph is made up of a fixed cost c_0, a cost c_1 per sampling unit and a cost c_2 per observation unit. Then the sampling of n nodes from an undirected graph with N nodes and the observation of the associated subgraph will cost

$$c_0 + n\,c_1 + \binom{n}{2} c_2. \tag{16}$$

If the associated partial graph is observed, the total cost will be

$$c_0 + n\,c_1 + n(2N-n-1)c_2/2. \tag{17}$$

If the purpose of the investigation is to get an estimate of R with a prescribed value of the variance, the variance formulas together with the cost functions (16) and (17) will be the basic means of choosing the cheapest observation procedure.

3. ESTIMATION OF THE NODE AND ARC TOTALS

3.1 Introduction

In Section 4 of the previous chapter the population graph was generalized from a simple graph to an undirected graph with node and arc values. Moreover, the estimation problem was generalized from the estimation of the arc frequency to the estimation of the node and arc totals. Sampling without replacement was incorporated in a more general sampling scheme. The same generalizations will now be considered here, and the notations that were introduced in Chapter 4, Subsections 4.1 and 4.2 will be used.

The sample graph is a partial graph of the population graph obtained by node sampling. The node values of the nodes in the sample are observed. The arc values of the arcs that have at least one of their nodes in the sample are also observed. Thus, the sample graph consists of all the node and arc values that do not belong to the subgraph associated with the complement of the node sample.

By use of the symmetric matrix C which consists of the node and arc values of the population graph, the sample information can be described as the elements of C that belong to any of the sampled rows (columns).

3.2 Unbiased estimators and their variances

The problem is to estimate the node and arc totals T_1 and T_2 defined according to (48) in Chapter 4. The node and arc totals of the sample graph can be written

$$t_1 = \sum_{i=1}^{N} C_{ii}\,\theta_i \tag{18}$$

and

$$t_2 = \sum_{i<j} \sum C_{ij}(\theta_i+\theta_j-\theta_i\theta_j). \tag{19}$$

Their expected values are immediately obtained as $E\,t_1 = T_1 p_1$, and $E\,t_2 = T_2(2p_1-p_2)$. Hence it follows that T_1 and T_2 have the unbiased estimators

$\hat{T}_1 = t_1/p_1$ and $\hat{T}_2 = t_2/(2p_1-p_2)$. The estimator $\hat{T}_1$ is identical with the corresponding estimator of Chapter 4, and $\sigma_1^2 = \text{Var } \hat{T}_1$ is given by (60) in Chapter 4. It is reproduced here for ease of reference:

$$\sigma_1^2 = \text{Var } \hat{T}_1 = T_{11}(1/p_1-1) + T_{12}(p_2/p_1^2-1) \ . \tag{20}$$

In order to obtain the variance of the estimator $\hat{T}_2$, the variable t_2 is first rewritten as

$$t_2 = \sum_{i<j}\sum C_{ij}[1-(1-\theta_i)(1-\theta_j)] = T_2-t_2' \tag{21}$$

where t_2' is the arc total of the subgraph associated with the complement of the sample. As a consequence, the variance of t_2 can be obtained by using the results of the previous chapter. Put for different indices i, j, k, l

$$\left.\begin{aligned} q_1 &= E\,(1-\theta_i) = 1-p_1 \ , \\ q_2 &= E\,(1-\theta_i)(1-\theta_j) = 1-2p_1+p_2 \ , \\ q_3 &= E\,(1-\theta_i)(1-\theta_j)(1-\theta_k) = 1-3p_1+3p_2-p_3 \ , \\ q_4 &= E\,(1-\theta_i)(1-\theta_j)(1-\theta_k)(1-\theta_l) = 1-4p_1+6p_2-4p_3+p_4 \ . \end{aligned}\right\} \tag{22}$$

Thus, q_v is the probability that not one of v specified nodes is included in the sample. Now

$$\sigma_2^2 = \text{Var } \hat{T}_2 = [q_2^2/(1-q_2)^2]\ \text{Var}\ (t_2'/q_2) \ , \tag{23}$$

and it follows from (60) in Chapter 4 (with p_v replaced by q_v) that

$$\sigma_2^2 = T_{22}\, q_2/(1-q_2) + T_{23}(q_3-q_2^2)/(1-q_2)^2 + T_{24}(q_4-q_2^2)/(1-q_2^2) \ . \tag{24}$$

To compute the covariance $\sigma_{12} = \text{Cov}\,(\hat{T}_1, \hat{T}_2)$, t_2 is written as in (21), and analogously $t_1 = T_1-t_1'$, where t_1' is the node total of the subgraph associated with the complement of the sample. Then

$$\sigma_{12} = [q_1 q_2/(1-q_1)(1-q_2)] \text{ Cov } (t'_1/q_1, t'_2/q_2) , \tag{25}$$

and according to (60) in Chapter 4 it follows that

$$\sigma_{12} = T_{32} q_2/(1-q_2) + T_{33} (q_3-q_1 q_2)/(1-q_1)(1-q_2) . \tag{26}$$

3.3 Unbiased estimators of the variances

The variances σ_1^2 and σ_2^2 and the covariance σ_{12} of the two estimators $\hat{T}_1$ and $\hat{T}_2$ can be unbiasedly estimated by substitution of unbiased estimators of the subgraph totals T_{uv} appearing in the expressions (20), (24), and (26).

Since it is assumed that it is known which nodes of the sample graph belong to the sample and which to the complement of the sample, the subgraph totals T_{uv} can be estimated in exactly the same way as in Chapter 4 by using only that part of the information that belongs to the subgraph associated with the sample. However, it is easy to modify these estimators by using more of the information provided by the sample graph.

The subgraph totals T_{11}, T_{12}, T_{32}, and T_{33} are estimated exactly as in Chapter 4. Consider then the subgraph total T_{22}. The corresponding sample graph parameter is

$$t_{22} = \sum_{i<j} \sum C_{ij}^2 (\theta_i + \theta_j - \theta_i \theta_j) , \tag{27}$$

with the expected value $E\, t_{22} = T_{22}(2p_1 - p_2)$. Thus an unbiased estimator of T_{22} is $\hat{T}_{22} = t_{22}/(2p_1 - p_2) = t_{22}/(1-q_2)$.

The subgraph total T_{23} is the sum of certain values of the subgraphs consisting of three nodes. These subgraphs can be separated into four categories, i.e. the subgraphs with three, two, one, or no nodes belonging to the sample. The subgraph value can be observed in the sample graph only for those subgraphs that have at least two nodes belonging to the sample. The subgraph values of the other subgraphs cannot be computed from the information provided by the sample graph. It follows that the subgraph total t_{23} of the sample graph can be written

$$t_{23} = \sum_{i<j<k} \sum \sum 2(C_{ij}C_{ik}+C_{ij}C_{jk}+C_{ik}C_{jk}) [\theta_i\theta_j\theta_k+\theta_i\theta_j(1-\theta_k) +$$

$$+ \theta_i(1-\theta_j)\theta_k+(1-\theta_i)\theta_j\theta_k] . \qquad (28)$$

Hence, the expected value becomes equal to $E\, t_{23} = T_{23}(3p_2-2p_3)$, and an unbiased estimator of T_{23} is $\hat{T}_{23} = t_{23}/(3p_2-2p_3) = t_{23}/(1-3q_2+2q_3)$.

Consider finally T_{24}. The position here is that the sample graph contains information about the values of those subgraphs with four nodes that are characterized by the fact that at least three of the nodes belong to the sample. Consequently, the subgraph total t_{24} of the sample graph can be written

$$t_{24} = 2 \sum_{i<j<k<l} \sum \sum \sum (C_{ij}C_{kl}+C_{ik}C_{jl}+C_{il}C_{jk}) [\theta_i\theta_j\theta_k\theta_l+\theta_i\theta_j\theta_k(1-\theta_l) +$$

$$+ \theta_i\theta_j(1-\theta_k)\theta_l+\theta_i(1-\theta_j)\theta_k\theta_l+(1-\theta_i)\theta_j\theta_k\theta_l] . \qquad (29)$$

The expected value becomes equal to $E\, t_{24} = T_{24}(4p_3-3p_4)$, and an unbiased estimator of T_{24} is $\hat{T}_{24} = t_{24}/(4p_3-3p_4) = t_{24}/(1-6q_2+8q_3-3q_4)$.

4. ESTIMATION OF THE DISTRIBUTION OF THE LOCAL DEGREES

4.1 The undirected case

The estimation problem that was considered in Chapter 4, Subsection 5.2 will be taken up here, the sample graph being a partial graph obtained by node sampling. As in the previous subsections of the present chapter, it is assumed that it is known which nodes in the observed partial graph belong to the sample. This is an important assumption, which implies that one knows which local degrees of the sample graph are also valid in the population graph.

The local degrees of the nodes of the sample are given by a vector x with the components $x_1, x_2, \ldots, x_n$. These local degrees constitute a random sample chosen without replacement from the local degrees $a_1, a_2, \ldots, a_N$ of the population graph. It is therefore possible to estimate in the ordinary way the proportion $p_a(u)$ of nodes of the population that have the local degree equal to u. The corresponding proportion $p_x(u)$ of the sample is an unbiased estimator of $p_a(u)$

with the variance

$$\sigma^2 = \text{Var } p_x(u) = (N-n)p_a(u)\,[1-p_a(u)]/(N-1)n. \tag{30}$$

An unbiased estimator of this variance is obtained in the ordinary way. The estimator becomes equal to

$$\hat{\sigma}^2 = (N-n)p_x(u)\,[1-p_x(u)]/N(n-1). \tag{31}$$

4.2 The directed case

If the population graph is directed, one has to distinguish between the case where only out-arcs are observed at the nodes of the sample and the one where both out-arcs and in-arcs are observed.

If both out-arcs and in-arcs are observed, it is possible to proceed as in the previous subsection. Denote the out-degrees by $x_1, x_2, \ldots, x_n$ and the in-degrees by $y_1, y_2, \ldots, y_n$ at the sample nodes. The out-degrees and the in-degrees of the population graph are denoted as before by $a_1, a_2, \ldots, a_N$ and $b_1, b_2, \ldots, b_N$, respectively. Then it is evident that the pairs of numbers $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ may be considered as a random sample chosen without replacement from the set of pairs $(a_1, b_1), (a_2, b_2), \ldots, (a_N, b_N)$. The proportion $p_{a,b}(u,v)$ of nodes with out-degree u and in-degree v is unbiasedly estimated by the corresponding proportion $p_{x,y}(u,v)$ in the sample. The variance and its unbiased estimator are obtained in the standard way for sampling without replacement.

If only the out-arcs are observed, the proportion $p_a(u)$ can be unbiasedly estimated by $p_x(u)$, and the variance and its estimator are obtained in the standard way. A more difficult problem is to estimate the distribution of the in-degrees by use of the observed out-arcs.

Let $y_1, y_2, \ldots, y_N$ denote the in-degrees of the sample graph. Note that not only the sample nodes but all the nodes of the sample graph are considered. It holds that

$$y_j = \sum_{i=1}^{N} C_{ij}\,\theta_i \tag{32}$$

for $j = 1,2,\ldots,N$. Consider the simplest case of a population graph without circuit arcs and multiple arcs. Then the number $f_y(v)$ of nodes with the in-degree v in the sample graph has the expected value

$$E\,f_y(v) = \sum_{j=1}^{N} \binom{b_j}{v}\binom{N-b_j}{n-v} \Big/ \binom{N}{n} . \tag{33}$$

This expected value can be written

$$E\,f_y(v) = \sum_{V=0}^{N-1} P(v,V) f_b(V) , \tag{34}$$

with

$$P(v,V) = \binom{V}{v}\binom{N-V}{n-v}/\binom{N}{n}, \tag{35}$$

for $v = 0,1,\ldots,n$ and $V = 0,1,\ldots,N-1$. The equation system (34) consists of $n+1$ equations and has N unknown frequencies $f_b(V)$. In order to obtain a unique solution when $n+1<N$, it is possible to use methods similar to those applied in Chapter 4, Sections 5 and 6. If it is assumed that $f_b(V) = 0$ for $V>n$, the solution of the equation system (34) is obtained as

$$f_b(V) = \sum_{v=0}^{n} Q(V,v)\,E\,f_y(v) \tag{36}$$

for $V = 0,1,\ldots,n$, with

$$Q(V,v) = \binom{v}{V}\binom{n-v}{N-V}/\binom{n}{N} = \binom{N}{V}\binom{n-N}{v-V}/\binom{n}{v} \tag{37}$$

for $V = 0,1,\ldots,n$ and $v = 0,1,\ldots,n$. If the observed values $f_y(v)$ are substituted for the expected values $E\,f_y(v)$ in (36), one obtains unbiased estimators $\hat{f}_b(V)$ of $f_b(V)$ for $V = 0,1,\ldots,n$.

The variances and the covariances of the estimators $\hat{f}_b(V)$ can be given as

$$\operatorname{Cov}\left[\hat{f}_b(U), \hat{f}_b(V)\right] = \sum_{u=0}^{n} \sum_{v=0}^{n} Q(U,u)\, Q(V,v)\, \operatorname{Cov}\left[f_y(u), f_y(v)\right]. \tag{38}$$

By a combinatorial argument it can further be proved that

$$E\, f_y(u) f_y(v) = \sum_{i=1}^{N} \sum_{j=1}^{N} P(y_i=u, y_j=v) = I(u=v)\, E\, f_y(v) +$$

$$+ \sum_{i\neq j} \sum_{w=0}^{B_{ij}} \binom{B_{ij}}{w} \binom{b_i - B_{ij}}{u-w} \binom{b_j - B_{ij}}{v-w} \binom{N-b_i-b_j+B_{ij}}{n-u-v+w} \Big/ \binom{N}{n}, \tag{39}$$

with

$$B_{ij} = \sum_{k=1}^{N} C_{ki}\, C_{kj}. \tag{40}$$

To be able to proceed any further in this way, one has to introduce the simultaneous distribution of (B_{ij}, b_i, b_j) and to estimate its frequencies by using the sample graph. As in Subsection 5.3 of Chapter 4, these methods are too involved to be of any practical use.

4.3 A numerical example

To illustrate the use of the formulas given above, a simple example will be considered.

Let the population graph be the graph of Figure 2. It has the arc frequency matrix

	1	2	3	4	5
1	0	1	0	1	0
2	0	0	1	0	0
3	0	1	0	0	0
4	0	0	1	0	0
5	0	0	0	0	0

and the in-degrees are 0, 2, 2, 1, 0, i.e. $f_b(0)=2$, $f_b(1)=1$, and $f_b(2)=2$. There are

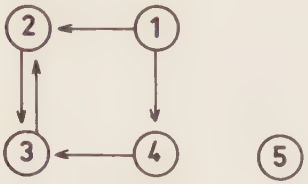


Figure 2

Table 3. Distribution of the in-degrees associated with the partial graphs obtained by sampling three nodes from the population graph of Figure 2.

Sample	$f_y(v)$ for v= 0, 1, 2, 3
1 2 3	2 2 1 0
1 2 4	2 2 1 0
1 2 5	2 3 0 0
1 3 4	2 2 1 0
1 3 5	3 1 1 0
1 4 5	2 3 0 0
2 3 4	3 1 1 0
2 3 5	3 2 0 0
2 4 5	4 0 1 0
3 4 5	3 2 0 0

ten samples of n=3 nodes. The distributions of the in-degrees of the sample graphs are listed in Table 3. Suppose that $f_b(4)=0$ is known and that $f_b(V)$ is to be estimated for V = 0,1,2,3. The numbers Q(V,v) are given by the elements of the matrix

$$Q = \begin{bmatrix} 1 & -2/3 & 1 & -4 \\ 0 & 5/3 & -10/3 & 15 \\ 0 & 0 & 10/3 & -20 \\ 0 & 0 & 0 & 10 \end{bmatrix} \tag{41}$$

and the values of the estimators

$$\hat{f}_b(V) = \sum_{v=0}^{n} Q(V,v) f_y(v) \tag{42}$$

are given in Table 4. The sampling distribution of the estimators is given in Table 5. A simple check shows that the estimators are unbiased, i.e. $E\,\hat{f}_b(V) = 2,1,2,0$ for $V = 0,1,2,3$, respectively. Moreover, it is found that $\mathrm{Var}\,\hat{f}_b(V) = 19/9,\ 71/9,\ 8/3,\ 0$ for $V = 0,1,2,3$, respectively.

Table 4. The estimators $\hat{f}_b(V)$ corresponding to the sample size 3 and the population graph shown in Figure 2.

Sample	$3\,\hat{f}_b(V)/5$ for V = 0, 1, 2, 3
1 2 3	1 0 2 0
1 2 4	1 0 2 0
1 2 5	0 3 0 0
1 3 4	1 0 2 0
1 3 5	2 -1 2 0
1 4 5	0 3 0 0
2 3 4	2 -1 2 0
2 3 5	1 2 0 0
2 4 5	3 -2 2 0
3 4 5	1 2 0 0

Table 5. The sampling distribution of the estimators $\hat{f}_b(V)$ corresponding to the sample size 3 and the population graph shown in Figure 2.

$3\hat{f}_b(V)/5$ for V = 0, 1, 2, 3	Frequency
1 0 2 0	3
0 3 0 0	2
2 -1 2 0	2
1 2 0 0	2
3 -2 2 0	1

4.4 Comments about applications

Some fields of application of estimation by use of partial graph sampling will be indicated. The exposition is vague and is only intended to illustrate the use of graphs in various applications of interest.

Example 1. In populations of ants one may observe that the ants touch each other or seem to communicate. Such contacts between the ants can in principle be represented by a graph, but for large populations it is impossible in practice to construct the graph. However, it may be of value to know the graph in order to be able to answer such questions as: Do the contacts take place only between some of the ants or between all of them? Do some of the ants have more contacts than the others?

In populations of hares, elks or other animals, it is also possible to state problems about certain contact frequencies that may be of interest in animal behavior research.

If it is possible in any way to identify the animals between which the contacts are observed, graph methods may be applied. Suppose, for instance, that a sample of animals is chosen, and that these animals can be distinguished (for instance by different color marks). Moreover, suppose that each animal in the sample can be observed in such a way that one knows with which other animals

it has contacts (for instance by color marking those animals). Then the knowledge gathered in this way corresponds to a partial graph of the population graph. An estimation problem of interest may be to estimate the number of animals with the different contact frequencies, i.e. to estimate the distribution of the local degrees in the population graph.

Example 2. Suppose that the internal telephone contacts in an enterprise with N employees are to be studied. Particularly, it is desired to find out how often a person called is not available. A sample of n persons is chosen at random, and information is obtained from these persons about how often they have not been successful when ringing different colleagues. Introduce a graph with N nodes representing the employees, and put an arc from i towards j if i has called j in vain too often in any specified meaning. The information collected corresponds to the partial graph with all the out-arcs from the sampled nodes. An estimation problem of interest may be to estimate the number of persons who have been rung in vain too often by at least one of their colleagues. Translated to graph terminology, it is desired to estimate the number of nodes that have at least one in-arc. Expressed by matrix language, the problem is to estimate the number of positive column sums of the arc frequency matrix C. With the notations used above, the parameter to be estimated is $N-f_b(0)$.

5. ESTIMATION OF THE NUMBER OF CONNECTED COMPONENTS

5.1 Introduction

In Chapter 4, Section 6, the problem considered was that of estimating the number of connected components of the population graph by use of a sampled subgraph. When the sample graph is a partial graph, there is more information available about the population graph than in the subgraph sampling case. It does not appear that the estimation problem for a general population graph is any easier owing to the information increase. It is still possible for the sample graph to have either more or fewer components than the population graph. How-

ever, the problem becomes extremely simple in the particular case of a population graph with all its components complete.

5.2 Complete components

The notation that was introduced in Chapter 4, Subsection 6.1 will also be used here. When all the components of the population graph are complete, the sample graph provides information not only about which nodes of the sample belong to the same components of the population graph, but also the reach in the population graph of every sample node. According to (103) in Chapter 4, the number K of components can be written as the sum of the inverted reaches U_i. It is therefore possible to estimate K by the corresponding sum for the sample nodes, after a proper correction has been made to ensure that there is no bias. In fact,

$$\hat{K} = \sum_{i=1}^{N} \theta_i / p_1 U_i \tag{43}$$

is immediately seen to be unbiased, and routine calculations show that

$$\sigma^2 = \operatorname{Var} \hat{K} = (1/p_1 - p_2/p_1^2) \sum_{i=1}^{N} 1/U_i^2 + (p_2/p_1^2 - 1)K^2 , \tag{44}$$

and it is readily found that σ^2 has the unbiased estimator

$$\hat{\sigma}^2 = (1/p_2 - 1/p_1) \sum_{i=1}^{N} \theta_i / U_i^2 + (1/p_1^2 - 1/p_2)\Big(\sum_{i=1}^{N} \theta_i / U_i \Big)^2 . \tag{45}$$

The formulas (44) and (45) are completely analogous to the formulas (63) and (66) for σ_1^2 and $\hat{\sigma}_1^2$ in Chapter 4, if c in (63) is interpreted as the vector of the inverted reaches in the population, and z in (66) is interpreted as the vector of the inverted reaches in the sample.

Chapter 6

INFERENCE FROM FLOW MEASUREMENTS

SYNOPSIS

1. Introduction. Applications are given of flow estimation in graphs from operations research and statistics. Especially two examples are discussed which refer to log floating and volumes of mail.

2. A model of flow measurement. A model of flow measurement is given. The flow estimation problem is formulated and discussed. The problem is reduced to a simple but general form.

3. Application of linear theory. The theory of linear spaces is applied to the flow estimation problem. A consequential use is made of the matrix concepts range space, null space, and generalized inverse. In this way an elegant and compact account is achieved. Admissible, consistent, and unique flows are studied. The results are interpreted by application of both algebra and graph theory. The Gauss-Markov estimation theory is applied to the least squares estimators of the arc flows.

4. Flow estimation in complete graphs. The flow estimation problem is explicitly solved for complete graphs. An application to a class transition problem is given.

5. Flow estimation in complete bipartite graphs. Explicit solutions are given. The results are applied to two-way contingency tables.

6. Flow estimation in trees. The concept of terminal node class is defined for trees. An algorithmic method of solution of the flow estimation problem is based on this concept. An explicit solution is given for regular trees. The simple interpretation which is possible in this case is shown not to be general. An application to nested classifications is given.

1. INTRODUCTION

1.1 Flows in graphs

In operations research flows in graphs may be used as models of phenomena concerning real or fictitious transports of units. A few examples are transportation networks for distribution of goods, allocation of resources to different activities, communications systems for information transmission, money transactions between enterprises, assignment of units to different classes.

The literature about flows in graphs deals with various types of optimization problem, i.e. finding optimal routes and determining maximal flows. The methods of solution consists of special flow algorithms and general linear programming algorithms. There are several good accounts of the theory. Gale (1960), Ford and Fulkerson (1962), and Berge and Ghouila-Houri (1965) may be mentioned as a few examples.

When a flow system is to be inspected or controlled, it may be appropriate to measure or to estimate the arc flows of the graph. The estimation problem of this chapter is concerned with the question of how the incomplete and uncertain information about the flow that is available may be used to choose optimal estimators of the arc flows. If there is a possibility of making a choice as to how the information about the flow is to be collected, for instance by allocation of measurements in different parts of the graph, then there is also a design problem involved.

Besides the operations research applications of flows in graphs, the applicability to statistical and demographical population estimation problems must also be noted. Consider a population of units which are associated with a numerical characteristic. If the totals of different subpopulations, disjoint or overlapping, are independently estimated, the estimators may be represented as the observed arc flows in a graph showing the connections between the various subpopulations. There will generally be a lack of consistency between the estimators and the interrelations of the subpopulations. There is then a need of a procedure for obtaining consistent estimators of the subpopulation totals.

A few examples will now be sketched to illustrate some flow estimation situations. Some other applications are provided in Sections 4, 5, and 6 below.

1.2 Log floating

Consider a river system for log floating, with several input places and several inspection places, at which latter the amount of some specified sort of timber is estimated. In Figure 1 is shown a simple river system with eight inspection places. At four of these (3, 6, 7, and 8), the input amounts are observed, and at the other inspection places the amounts passing by are observed. How can these observations be combined to an estimator of the total amount of this sort of timber? This problem may be formulated as a flow estimation problem in a tree graph. Figure 2 shows a tree graph corresponding to the river system. The nodes correspond to the inspection places. Imagine that timber flows through the arcs. The terminal nodes 3, 6, 7, and 8 are input places. In the non-terminal nodes, the incoming flow equals the outgoing flow. The real arc flows are unknown. The arc flow observations obtained need not satisfy the condition of equality between incoming and outgoing flows. The problem is to combine the arc flow observations in such a way as to get good estimators of the arc flows.

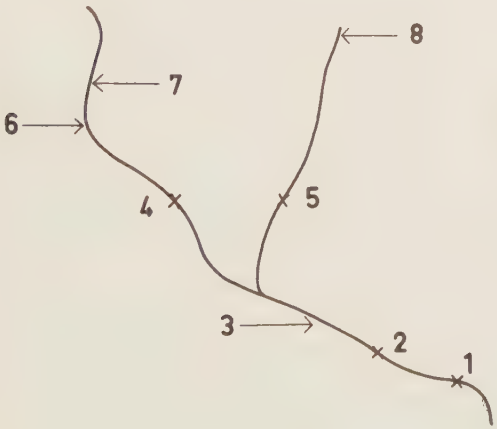


Figure 1

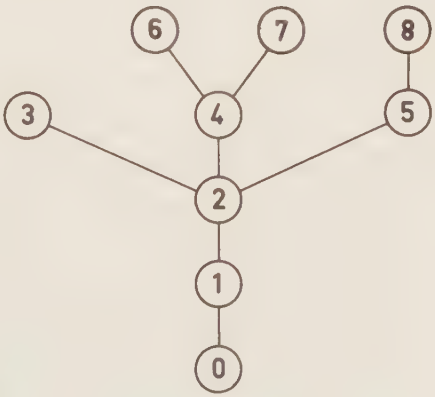


Figure 2

1.3 Mail volumes

Consider the outgoing or incoming volume of mail at a certain post office. By tracing the mail forward or backward, via successive sorting and distribution at intermediate post offices, to the destinations or starting-points, one gets a tree with mail flow. The volumes of incoming and outgoing mail may be estimated on some sort of sampling basis.

Severo and Newman (1960) have described a sampling and estimation procedure which has been applied to estimating the volumes of outgoing mail to different places in per cent of the volume of incoming mail. Independent estimators are given of the relative outflows at the nodes. The product of the estimators encountered from the root to a terminal node is an unbiased estimator of the relative volume of incoming mail to the terminal node. The variance of the estimator is easily obtained from the means and variances of the estimators at the nodes. This is shown and applied in the paper by Severo and Newman referred to above.

2. A MODEL OF FLOW MEASUREMENT

2.1 The general problem

In general terms, the flow estimation problem can be formulated as follows. There is given a directed graph with arc flows in the directions of the arcs. The graph may have multiple and opposite directed arcs between the nodes. In some of the nodes flow may be generated, and in some of the nodes flow may be absorbed. A flow generating node has a positive net outflow, and a flow absorbing node has a negative net outflow. At any other node where the flow passes through, the inflow equals the outflow, i.e. the net outflow is zero. The net outflows of all the nodes add up to zero.

There is incomplete knowledge about the flows of the graph. Some arc flows may be known and some unknown. The unknown arc flows are inaccurately determined by measurement. Some nodes may have known net outflows, others may have unknown net outflows. The sizes of the flows and the existing informa-

tion about the structure of the graph are to be combined to consistent and best possible (properly understood) estimators of the arc flows.

2.2 Notations

Let the graph have a node set Ω with n nodes. Let M_{ij} denote the set of arcs going from node i towards node j. Let A_i and B_i denote the set of arcs after and before node i, i.e. $A_i = \bigcup_{j \in \Omega} M_{ij}$ and $B_i = \bigcup_{j \in \Omega} M_{ji}$ for $i \in \Omega$. The set of all the arcs is denoted M.

Associated with each arc $\alpha \in M$, there is an arc flow u_α, and associated with each node $i \in \Omega$ a net outflow $a_i = u(A_i) - u(B_i)$, where $u(S) = \sum_{\alpha \in S} u_\alpha$ for arbitrary sets S. Analogous notation will be used for functions other than u. The net outflows will also be called node flows. From the definitions it follows that the sum of the node flows equals zero.

The node flows a_i are assumed to be known for the nodes $i \in \omega$, where ω is a subset of Ω containing $m \leq n$ nodes. If $m < n$, the node flows are unknown for the nodes $i \in \Omega - \omega$. In order to simplify the terminology, the nodes with known and unknown node flows will be called white and black, respectively. They may also be pictured in this way in figures, white and black nodes being represented by unfilled and filled-up circles, respectively. The known node flows may be written in the unfilled circles.

With each arc $\alpha \in M$ there is associated a measurement value x_α of the arc flow u_α. These measurements are assumed to be independent stochastic variables with the expected values $E\, x_\alpha = u_\alpha$ and the finite variances $\mathrm{Var}\, x_\alpha = \sigma_\alpha^2 \geq 0$. A zero variance corresponds to a known arc flow, and a positive variance corresponds to an arc flow that is inaccurately determined by a measurement. The variances will be assumed to be of known relative sizes, i.e. $\sigma_\alpha^2 = \sigma^2 D_\alpha$, where D_α is known for $\alpha \in M$, and σ^2 is a known or unknown proportional factor.

2.3 Least squares estimators

The measurements x_α for $\alpha \in M$ will not in general satisfy the conditions of consistency implied by the white nodes, i.e. $x(A_i) - x(B_i)$ and a_i are not, in general, equal for $i \in \omega$. The flow estimation problem is to use the measurements x_α for $\alpha \in M$ in order to get estimators of the arc flows u_α for $\alpha \in M$ that are consistent with the node flows a_i for $i \in \omega$.

According to a least squares criterion, the arc flows are estimated by the values u_α^* that minimize

$$\Sigma (x_\alpha - u_\alpha)^2/\sigma_\alpha^2 , \tag{1}$$

(where the sum is to be extended over the arcs with a positive measurement variance) subject to the side conditions $u(A_i) - u(B_i) = a_i$ for $i \in \omega$. If Lagrangian multipliers λ_i for $i \in \omega$ are introduced, and if

$$\Sigma (x_\alpha - u_\alpha)^2/\sigma_\alpha^2 + 2 \sum_{i \in \omega} \lambda_i [u(A_i) - u(B_i)] \tag{2}$$

is minimized, then

$$u_\alpha^* = x_\alpha + \sigma_\alpha^2 (\lambda_j - \lambda_i) \tag{3}$$

for $\alpha \in M_{ij}$. Here λ_i has to be interpreted as zero for $i \in \Omega - \omega$, and has for $i \in \omega$ to be determined from the equation system implied by the side conditions, i.e. from

$$x(A_i) + \sum_{j \in \omega} \lambda_j \sigma^2(M_{ij}) - \lambda_i \sigma^2(A_i) - x(B_i) - \lambda_i \sigma^2(B_i) + \\ + \sum_{j \in \omega} \lambda_j \sigma^2(M_{ji}) = a_i \tag{4}$$

for $i \in \omega$. Put

$$z_i = x(A_i) - x(B_i) - a_i \tag{5}$$

for $i \in \omega$, and

$$\sigma_{ij}^2 = \sigma^2(M_{ij}) + \sigma^2(M_{ji}) \tag{6}$$

for $i \neq j$, $i \in \Omega$, and $j \in \Omega$. Then the equation system (4) can be written as

$$z_i = \lambda_i \sum_{j \in \Omega} \sigma_{ij}^2 - \sum_{j \in \omega} \sigma_{ij}^2 \lambda_j \tag{7}$$

for $i \in \omega$.

It is possible to draw some interesting conclusions from the expression of the solution that is provided by (3) and (7). According to (7), the Lagrangian multipliers depend upon the arc and node flows only through z_i for $i \in \omega$, i.e. the observed minus the known node flows at the white nodes. This dependence can be completely given with recourse to σ_{ij}^2 for $i \neq j$, $i \in \Omega$, and $j \in \Omega$. It is possible to interpret σ_{ij}^2 as the variance of the observed net flow between node i and node j. According to (3), the estimator u_α^* depends upon the Lagrangian multipliers, the measurement x_α, and the variance σ_α^2 for the same arc $\alpha \in M$. It is therefore possible to separate the flow estimation problem into two parts. In the first part, the Lagrangian multipliers are determined according to (7), and only net flows between the nodes have to be considered. In the second part, the estimators are determined according to (3). As the problem is essentially related to the first part, it is no severe restriction only to consider net flow estimation.

2.4 Repeated measurements

So far it has been assumed that there is one single measurement for each unknown arc flow. If several independent measurements of the same arc flow are taken (not to be confused with multiple arcs), these measurements can be combined into a weighted mean that may be considered as a single measurement of the arc flow. If the weights are properly chosen, this substitution implies no restriction. This may be shown in the following way.

Let x_1 and x_2 be independent measurements of the same arc flow u.

Suppose that the measurement variances are σ_1^2 and σ_2^2, respectively. Then the sum of squares (1) contains

$$(x_1-u)^2/\sigma_1^2 + (x_2-u)^2/\sigma_2^2 \ , \tag{8}$$

which can be written

$$(x_1-x)^2/\sigma_1^2 + (x_2-x)^2/\sigma_2^2 + (x-u)^2/\sigma^2 \ , \tag{9}$$

if x and σ^2 are defined by

$$x = (x_1/\sigma_1^2 + x_2/\sigma_2^2)/(1/\sigma_1^2 + 1/\sigma_2^2) \ , \tag{10}$$

$$\sigma^2 = 1/(1/\sigma_1^2 + 1/\sigma_2^2) \ . \tag{11}$$

Here x is recognized as the unbiased linear combination of x_1 and x_2 that has the least possible variance σ^2. Especially if σ_1^2 and σ_2^2 are equal, the minimum variance combination is the ordinary arithmetic mean $x = (x_1+x_2)/2$. From (9) it follows that it is sufficient to use x as a single measurement of u with the variance σ^2.

By application of (10) and (11), all repeated measurements can be reduced to single measurements of the arc flows. The assumption that each arc flow has exactly one measurement therefore imposes no restriction.

An important special case is when all the measurements have the same known or unknown variance σ^2, and n_{ij} measurements are taken of the arc flow u_{ij}. The mean x_{ij} of the measurements then has the variance $\sigma_{ij}^2 = \sigma^2/n_{ij}$. This is an example of known relative variances.

2.5 General assumptions

According to the discussion above, it is no essential restriction to limit the problem to single measurements of net flows between the nodes. Thus, the graph may be considered as an undirected graph with no circuit arcs and no

multiple arcs. It is also possible to make some other assumptions which do not limit the applicability. Possibilities will now be considered which may be of value when the flow estimation problem is studied for special types of graph.

If the graph consists of several connected components, it is obvious that they do not influence each other. Hence it is sufficient to deal with connected graphs. If known arc flows are permitted to be zero, it is no restriction to assume that the graph is complete.

The graph may be supposed either to have positive measurement variances at all the arcs, or to have zero node flows at all the white nodes. If it is desired to have positive measurement variances at all the arcs, the arcs with known arc flows are omitted. This may be done if the node flows are properly modified. If it is desired to have zero node flows at all the white nodes, this is achieved by adding an arc with a properly chosen arc flow at each white node with a non-zero node flow.

The arcs between black nodes may be omitted if they have positive measurement variances, because then their measurements provide isolated information, i.e. have no influence upon the other arc flows. Therefore, each arc with a positive measurement variance may be assumed to have at least one white node.

A terminal node is a node with the local degree equal to one. Each terminal node may be assumed to be black, because a known arc flow between a terminal node and a non-terminal node may be omitted if the node flow is properly modified at the non-terminal node.

It is also possible to assume that each black node is a terminal node, because the flow estimation problem is not changed if each black non-terminal node is changed to separate black terminal nodes at each one of the arcs of the original black node.

2.6 Comments

The model of flow measurement in a graph that has been presented above has been inspired by a military flow control system which has been studied by the author at the Swedish Research Institute of National Defence.

The flow estimation problem that was first considered was concerned with tree graphs. In a paper by the author (Frank, 1969 b), an algorithm was introduced for estimating flows in tree graphs. The algorithm was formulated on the basis of the formula (10). By using matrix algebra, the algorithm was proved to give the least squares estimators. Borenius (1969) provided an alternative proof also by using matrix algebra. In Section 6 of this chapter the flow estimation problem for tree graphs will be taken up. There the algorithm will be derived in a much more elegant way by starting from the equation system (7).

The flow estimation in other graphs than trees has been dealt with in a paper by the author (Frank, 1969 c). The results of that paper are incorporated in Sections 4 and 5 of this chapter.

3. APPLICATION OF LINEAR THEORY

3.1 Preliminaries

Shilov (1961) and Graybill (1969) give good accounts of the theory of linear spaces appropriate for the present purpose. Some well-known results from the theory of linear spaces which will be needed in the following will be given in this section. By a consequential use of the null spaces and the range spaces of matrices, the results have been stated in a compressed and uniform style.

If A is an arbitrary matrix, R(A) denotes the range space (the column space) of A, i.e. the set of vectors y that can be written $y = Ax$ for some vector x.

If A is an arbitrary matrix, N(A) denotes the null space of A, i.e. the set of vectors x that satisfy $Ax = 0$.

If I denotes the unit matrix and $\perp$ means "orthogonal to", the following relations obtain:

$$R(A) \perp N(A') \ , \tag{12}$$

$$R(A) = N(I-AB) \text{ if B satisfies } ABA = A \ , \tag{13}$$

$$N(A) = R(I-BA) \text{ if B satisfies } ABA = A \ . \tag{14}$$

To prove (12) is must be shown that $y'z = 0$ if $y \in R(A)$ and $z \in N(A')$. There is a vector x such that $y = Ax$, and it follows that $y'z = x'A'z$, which equals zero as $A'z = 0$.

To prove (13), consider a vector $y \in N(I-AB)$. From $(I-AB)y = 0$ it follows that $y = ABy$, i.e. $y \in R(A)$. Consequently, $N(I-AB) \subset R(A)$ for arbitrary A and B. Then consider a vector $y \in R(A)$. There is a vector x such that $y = Ax$. If B is a matrix that satisfies $ABA = A$, it is possible to write $y = ABAx$. Substituting y for Ax gives $y = ABy$, i.e. $(I-AB)y = 0$. Thus $R(A) \subset N(I-AB)$ if B satisfies $ABA = A$. This implies that (13) applies.

It is possible to prove (14) analogously to (13). Let $x \in N(A)$, i.e. $Ax = 0$. For an arbitrary matrix B it follows that $BAx = 0$, $x = (I-BA)x$, i.e. $x \in R(I-BA)$. Hence $N(A) \subset R(I-BA)$ for arbitrary B. Then consider a vector $x \in R(I-BA)$. There is a vector y such that $x = (I-BA)y$. It follows that $Ax = (A-ABA)y$, which equals zero if $ABA = A$. Thus, $R(I-BA) \subset N(A)$ if $ABA = A$, and (14) follows.

Penrose (1955 and 1956) has shown that for an arbitrary matrix A there is a unique matrix B that satisfies $ABA = A$, $BAB = B$, $(AB)' = AB$, and $(BA)' = BA$. He calls B the generalized inverse of A. A matrix B that satisfies $ABA = A$ is called a conditional inverse of A (Graybill, 1969). The generalized inverse is a conditional inverse, but there may be other conditional inverses also.

Put $\| x \|^2 = x'x$. Then $\| b-y \| \geq \| b-y_0 \|$ for every y that belongs to a linear subspace L if, and only if, y_0 satisfies $y_0 \in L$ and $b-y_0 \perp L$. The vector y_0 which satisfies the conditions is, in geometric language, called the orthogonal projection of b upon L. It is denoted $y_0 = \text{Proj}(b|L)$.

The following expressions for the projections are obtained when the linear space L is specified as a range space or a null space:

$$\text{Proj}(b|R(A)) = ABb \text{ if } ABA = A \text{ and } (AB)' = AB \,, \tag{15}$$

$$\text{Proj}(b|N(A)) = (I-BA)b \text{ if } ABA = A \text{ and } (BA)' = BA \,. \tag{16}$$

To prove (15) it must be shown that $y_0 = ABb$ satisfies $y_0 \in L$ and $b-y_0 \perp L$ when $L = R(A)$. It is immediately clear that $ABb \in R(A)$. Moreover, according to (14) and (12), $b-ABb = (I-AB)b = (I-B'A')b \in R(I-B'A') = N(A') \perp R(A)$.

Analogous reasoning may be used to prove (16). It is also possible to use the relation $\text{Proj}(b|N(A)) = b - \text{Proj}(b|R(A'))$ and apply (15).

The projections as given by (15) and (16) may be calculated with the generalized inverse of A. However, it is generally more difficult to determine the generalized inverse than to get an arbitrary conditional inverse. Therefore it may be preferable to have recourse to such formulas for the projections which use an arbitrary conditional inverse. Iterative methods of computing conditional inverses have been given by Boot (1963), Pyle (1964), and Ben-Israel (1965). Further references are provided by Graybill (1969).

In order to get projection formulas involving conditional inverses, the case $L = R(A)$ is considered first. The condition $y_0 \in L$ implies that there is a vector x such that $y_0 = Ax$. According to (12), the condition $b - y_0 \perp L$ is equivalent to $b - y_0 \in N(A')$. Hence $A'(b - y_0) = 0$, i.e. $A'Ax = A'b$. Since there is a solution x to this equation, $A'b \in R(A'A)$. According to (13), $A'b \in N(I - A'AC)$, where C is an arbitrary conditional inverse of $A'A$, i.e. $A'ACA'A = A'A$. It follows that $(I - A'AC)A'b = 0$, i.e. $A'ACA'b = A'b$. Consequently, $x_0 = CA'b$ is a solution, and it is possible to write $y_0 = Ax_0 = ACA'b$, i.e.

$$\text{Proj}(b|R(A)) = ACA'b \text{ if } A'ACA'A = A'A \ . \tag{17}$$

For the other case $L = N(A)$ it follows from $y_0 \in L$ that $Ay_0 = 0$. The condition $b - y_0 \perp L$ implies, according to (12), that there is a vector x such that $b - y_0 = A'x$. Hence $AA'x = Ab$. The existence of a solution to this equation implies that $Ab \in R(AA')$. Then, according to (13), $Ab \in N(I - AA'C)$, where C is an arbitrary conditional inverse of AA', i.e. $AA'CAA' = AA'$. It follows that $(I - AA'C)Ab = 0$, i.e. $AA'CAb = Ab$. Consequently, $x_0 = CAb$ is a solution, and it is possible to write $y_0 = b - A'x_0 = (I - A'CA)b$, i.e.

$$\text{Proj}(b|N(A)) = (I - A'CA)b \text{ if } AA'CAA' = AA' \ . \tag{18}$$

3.2 Notations

Turning to the flow estimation problem, the following notations are introduced. The graph is connected and consists of a node set Ω with n nodes labeled $1,2,\ldots,n$. There are m white and n-m black nodes. The set of white nodes will be denoted ω, and the white nodes labeled $1,2,\ldots,m$. The graph has r arcs, and each arc has at least one white node. M denotes the set of the r pairs of nodes (j,k) with $j<k$ (for instance) such that there is an arc between node j and node k. The structure of the subgraph with the white nodes is given by an $m\times r$ matrix B. The matrix element corresponding to the row index i and the column index (j,k) is given by

$$B_{ijk} = \begin{cases} 1 & \text{if } i = j, \\ -1 & \text{if } i = k, \\ 0 & \text{otherwise}, \end{cases} \tag{19}$$

for $i = 1,2,\ldots,m$ and $(j,k)\in M$. There are r-n+m columns of B which correspond to the arcs between two white nodes. Each one of these columns consists of one digit 1, one digit -1 and m-2 zeros. There are n-m columns of B that correspond to the arcs with one white and one black node. Each one of these columns consists of one digit 1 and m-1 zeros.

The net flows from node j to node k for $(j,k)\in M$ are given by an $r\times 1$ vector u with the elements u_{jk} for $(j,k)\in M$. The corresponding measurements are given by an $r\times 1$ vector x with the elements x_{jk} for $(j,k)\in M$. The stochastic vector x has the expected value $E\,x = u$ and the covariance matrix $\mathrm{Cov}\,x = D$. D is an $r\times r$ diagonal matrix with the positive diagonal elements σ^2_{jk} for $(j,k)\in M$. The node flows of the white nodes are given by an $m\times 1$ vector a with the elements a_i for $i = 1,2,\ldots,m$.

It is practical to put $u_{jk} = -u_{kj}$ for $(k,j)\in M$, and $u_{jk} = 0$ if $(j,k)\notin M$ and $(k,j)\notin M$. Analogously, $x_{jk} = -x_{kj}$ for $(k,j)\in M$, and $x_{jk} = 0$ if $(j,k)\notin M$ and $(k,j)\notin M$. Also put $\sigma^2_{jk} = \sigma^2_{kj}$ for $(k,j)\in M$, and $\sigma^2_{jk} = 0$ if $(j,k)\notin M$ and $(k,j)\notin M$.

3.3 Admissible node flows

The graph structure represented by the matrix B and the node flows given by the vector a are the basic data of the flow estimation problem. However, it is not possible to specify the node flows quite arbitrarily. For instance, it is necessary in the case of graphs with no black nodes that the node flows add up to zero, i.e.

$$\sum_{i=1}^{n} a_i = 0 . \tag{20}$$

The possible node flows have to be obtainable by some arc flows. There must be some arc flows that satisfy

$$\sum_{j=1}^{n} u_{ij} = a_i \tag{21}$$

for $i = 1, 2, \ldots, m$. This condition of consistency between the arc flows and the node flows can be written more compactly as $Bu = a$. It follows that the admissible node flows are given by the vectors a which are linear combinations of the column vectors of B, i.e. $a \in R(B)$. It is customary to express the condition in the following way: The rank of B equals the rank of B extended with the column a.

In order to interpret the condition of admissibility in graph terms, the following correspondencies are noted. The column vectors of B correspond to the arcs of the graph. A set of column vectors is linearly independent if, and only if, it corresponds to a partial graph with no (undirected) circuits and with no connected black nodes. A basis of R(B) may be obtained, for instance, as a partial graph which consists of a tree for the white nodes together with one arc with a black node, if such an arc exists. When there are only white nodes in the graph, i.e. when $m = n$, it follows that rank $B = m-1 = n-1$. Hence the condition (20) is necessary and sufficient for the node flow vector a to be admissible when $m = n$. When the graph has at least one black node, i,e, when $m < n$, it follows that rank $B = m$. Consequently, all node flow vectors a are admissible when $m < n$.

3.4 Consistent flows

To each admissible node flow vector a, there is an arc flow vector u_0 that satisfies the condition $Bu_0 = a$. However, there may be other arc flow vectors u that satisfy $Bu = a$. Each vector u which satisfies $B(u-u_0) = 0$ is a solution to $Bu = a$, and every solution can be obtained in that way. Consequently, $u-u_0 \in N(B)$. To get an explicit formula of all the possible arc flow vectors u, the relation (14) is used. It follows that $u-u_0 \in R(I-AB)$ if A satisfies $BAB = B$. Each possible arc flow vector u can be written

$$u = u_0 + (I-AB)w \quad , \tag{22}$$

for some vector w, and every vector w gives a possible arc flow vector u through (22). Especially when $m < n$, rank $B = m$, and A may be chosen as $A = B'(BB')^{-1}$.

A possible arc flow vector u_0 can be determined in several ways. An algebraic way is to choose $u_0 = Aa$. It follows from $a \in R(B) = N(I-BA)$ for $BAB = B$ that $(I-BA)a = 0$, i.e. $BAa = a$, and consequently $Bu = a$ has a solution $u_0 = Aa$.

The graph may be used to get a possible solution u_0. Starting from an arbitrary partial tree of the graph, all the arc flows not belonging to the tree are put equal to zero. If there are black nodes, the arc flows are put equal to zero for all but one black node. The remaining tree with no more than one black node has its arc flows uniquely determined by the node flows of the white nodes.

3.5 Unique arc flows

A consequence of the previous results is that the equation $Bu = a$ with an admissible vector a has a unique solution u_0 if, and only if, the null space N(B) consists of the zero vector only. There are then r linearly independent row vectors of B, i.e. rank $B = r \leq m$. This implies that $B'B$ has an inverse, and the unique solution is given by $u_0 = (B'B)^{-1} B'a$.

How can the uniqueness be interpreted in the graph? As the graph is connected, $r \geq n-1$. It follows that uniqueness occurs only if $r = m-1 = n-1$, or if $r = m = n-1$. The first case means that the graph is a tree with no black nodes.

In the second case, the graph is a tree with one black node. It is easy to see that uniqueness is obtained by the node flows of the white nodes. Thus, the arc flows are unique if, and only if, the graph is a tree with no more than one black node.

3.6 An example

The determination of all the arc flows that are consistent with an arbitrary set of node flows will now be demonstrated for the simple graph of Figure 3.

The graph has $n = m = 5$ and $r = 8$. The node flows a_i for $i = 1, 2, \ldots, 5$ are admissible, i.e. $a_1 + a_2 + \ldots + a_5 = 0$. The matrix B becomes

	12	13	14	23	25	34	35	45
1	1	1	1	0	0	0	0	0
2	-1	0	0	1	1	0	0	0
3	0	-1	0	-1	0	1	1	0
4	0	0	-1	0	0	-1	0	1
5	0	0	0	0	-1	0	-1	-1

and rank $B = n-1 = 4$. A basis of R(B) is represented by a partial tree. In Figure 4 a partial tree is given together with the arc flows that are implied by the node flows. The corresponding conditional inverse A is given by

	1	2	3	4	5
12	0	-1	0	0	0
13	0	0	-1	0	-1
14	0	0	0	-1	0
23	0	0	0	0	0
25	0	0	0	0	0
34	0	0	0	0	0
35	0	0	0	0	-1
45	0	0	0	0	0

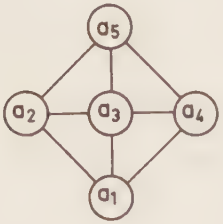


Figure 3

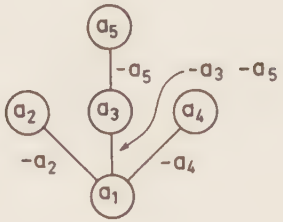


Figure 4

and the arc flow vector is $u_0 = Aa$. The matrix I-AB becomes

	12	13	14	23	25	34	35	45
12	0	0	0	1	1	0	0	0
13	0	0	0	-1	-1	1	0	-1
14	0	0	0	0	0	-1	0	1
23	0	0	0	1	0	0	0	0
25	0	0	0	0	1	0	0	0
34	0	0	0	0	0	1	0	0
35	0	0	0	0	-1	0	0	-1
45	0	0	0	0	0	0	0	1

and all the possible arc flow vectors u can be obtained according to (22) as $u = Aa + (I-AB)w$ for arbitrary vectors w, i.e.

$$\left.\begin{aligned}
u_{12} &= -a_2 + w_{23} + w_{25}\,, \\
u_{13} &= -a_3 - a_5 - w_{23} - w_{25} + w_{34} - w_{45}\,, \\
u_{14} &= -a_4 - w_{34} + w_{45}\,, \\
u_{23} &= w_{23}\,, \\
u_{25} &= w_{25}\,, \\
u_{34} &= w_{34}\,, \\
u_{35} &= -a_5 - w_{25} - w_{45}\,, \\
u_{45} &= w_{45}\,.
\end{aligned}\right\} \qquad (23)$$

Now suppose that a_1 and a_5 are unknown, i.e. $n=5$ and $m=3$. In this case,

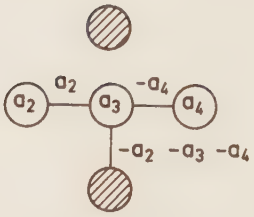


Figure 5

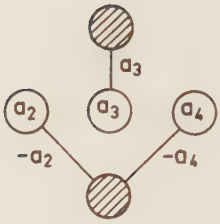


Figure 6

the matrix B becomes

	12	13	14	23	25	34	35	45
2	-1	0	0	1	1	0	0	0
3	0	-1	0	-1	0	1	1	0
4	0	0	-1	0	0	-1	0	1

and rank B = m = 3. A basis of R(B) may be represented by a partial graph which is a tree for the white nodes and has an arc to exactly one black node. Note, however, that there are other bases also. In Figures 5 and 6 two different bases and two consistent arc flows are shown. The matrix A corresponding to the basis of Figure 6 is

	2	3	4
12	-1	0	0
13	0	0	0
14	0	0	-1
23	0	0	0
25	0	0	0
34	0	0	0
35	0	1	0
45	0	0	0

and I-AB becomes

	12	13	14	23	25	34	35	45
12	0	0	0	1	1	0	0	0
13	0	1	0	0	0	0	0	0
14	0	0	0	0	0	-1	0	1
23	0	0	0	1	0	0	0	0
25	0	0	0	0	1	0	0	0
34	0	0	0	0	0	1	0	0
35	0	1	0	1	0	-1	0	0
45	0	0	0	0	0	0	0	1

It follows according to (22) that

$$\left.\begin{aligned} u_{12} &= -a_2 + w_{23} + w_{25}\,, \\ u_{13} &= w_{13}\,, \\ u_{14} &= -a_4 - w_{34} + w_{45}\,, \\ u_{23} &= w_{23}\,, \\ u_{25} &= w_{25}\,, \\ u_{34} &= w_{34}\,, \\ u_{35} &= a_3 + w_{13} + w_{23} - w_{34}\,, \\ u_{45} &= w_{45}\,. \end{aligned}\right\} \qquad (24)$$

3.7 Least squares estimators

When the arc flow vector u is not uniquely determined by the admissible node flow vector a, the observed arc flow vector x will be used to estimate u. According to a least squares criterion, u is estimated by the vector u^* that minimizes

$$(x-u)'\, D^{-1}\, (x-u) \qquad (25)$$

subject to the side condition Bu = a.

The covariance matrix D is a diagonal matrix with positive diagonal elements. If H is the diagonal matrix with elements that are the positive square roots of

the elements of D^{-1}, it is possible to write $D^{-1} = H^2 = H'H$. Let u_0 be an arbitrary flow vector that satisfies $Bu_0 = a$. Transform x and u to y and v according to $y = H(x-u_0)$ and $v = H(u-u_0)$. Then the problem becomes one of minimizing $\| y-v \|$ subject to the side condition $BH^{-1} v = 0$. The solution of this problem is

$$v^* = \text{Proj}(y|N(BH^{-1})) \ . \tag{26}$$

Application of (18) gives

$$v^* = (I-H^{-1}B'CBH^{-1})y \ , \tag{27}$$

where C is an arbitrary conditional inverse of $BH^{-1}(BH^{-1})' = BDB'$. Transformation backwards to the original variables gives

$$H(u^*-u_0) = (I-H^{-1}B'CBH^{-1})H(x-u_0) \ , \tag{28}$$

which simplifies to

$$u^* = x - DB'C(Bx-a) \ . \tag{29}$$

According to (29), the least squares estimators are generally determined with some arbitrary conditional inverse C of BDB'. When $m < n$, rank $B = m$. Then BDB' is non-singular with the inverse $C = (BDB')^{-1}$.

In the study of special graphs that follows in Sections 4, 5, and 6 it is preferable to have another expression for the solution. To get this expression, consider (26). Application of (12) to $y-v^* \perp N(BH^{-1})$ implies that $y-v^* = H^{-1}B'\lambda$ for some $m \times 1$ vector λ. Since $v^* \in N(BH^{-1})$, it follows that $BH^{-1}y = BDB'\lambda$. Returning to the original variables gives that λ is to be determined from

$$Bx - a = BDB'\lambda \ , \tag{30}$$

and then u^* is given by

$$u^* = x - DB'\lambda \ . \tag{31}$$

The $m \times 1$ vector $z = Bx - a$ has as elements z_i the observed minus the known node flow at node i for $i = 1, 2, \ldots, m$. The relation (30) can be written as the equation system

$$z_i = \lambda_i \sum_{j=1}^{n} \sigma_{ij}^2 - \sum_{j=1}^{m} \sigma_{ij}^2 \lambda_j \qquad (32)$$

for $i = 1, 2, \ldots, m$, i.e. the system (7). The equation system has at least one solution $\lambda_1, \lambda_2, \ldots, \lambda_m$ and exactly one solution if, and only if, $m < n$. Put $\lambda_i = 0$ for $i = m+1, m+2, \ldots, n$ and choose an arbitrary solution $\lambda_1, \lambda_2, \ldots \lambda_m$. Then the unique flow estimator is given according to (31) as

$$u_{ij}^* = x_{ij} + \sigma_{ij}^2 (\lambda_j - \lambda_i) \qquad (33)$$

for $(i,j) \in M$.

3.8 Properties of the estimators

In this subsection the expected value and the covariance matrix of the vector u^* of the least squares estimators will be derived, and the minimum value of the sum of squares (25) will also be determined. Any linear combination $b'u$ of the arc flows will be estimated from an unbiased linear combination $k+h'x$ of the observed arc flows. By application of the Gauss-Markov theory it will be shown that $b'u^*$ is the unique estimator of minimum variance in the class of unbiased linear estimators.

From (29) it follows directly that the least squares estimators are unbiased, i.e. $E\, u^* = u$. In order to determine the covariance matrix Cov u^*, the relation (27) and the fact that Cov $y = I$ are used. With the projection matrix

$$P = I - H^{-1}B'CBH^{-1}, \qquad (34)$$

the relation (27) becomes $v^* = Py$, and it follows that Cov $u^* = H^{-1}PIP'H^{-1}$. To get any further, use may be made of the rules of calculation with generalized inverses, which are found for instance in the book by Graybill (1969). However,

it is simpler to use (16) to prove that P is symmetric and idempotent, i.e. $P' = P$ and $P^2 = P$. Hence it follows that Cov $u^* = H^{-1}PH^{-1}$, and the substitution of (34) gives that

$$\text{Cov } u^* = D - DB'CBD \ . \tag{35}$$

Now turn to the minimum value of the sum of squares (25), i.e.

$$(x-u^*)' \ D^{-1} \ (x-u^*) \ . \tag{36}$$

By writing (36) as $\| y-v^* \|^2$ and using $v^* \perp y-v^*$, we have that (36) equals $y'(y-v^*) = y'(I-P)y$. Substitution of (34) implies that

$$(x-u^*)' \ D^{-1} \ (x-u^*) = (Bx-a)' \ C \ (Bx-a) \ . \tag{37}$$

This result may also be obtained from (29), but then the rules of calculation with generalized inverses are needed. It is possible to interpret (37) geometrically and to say that the distance between the observed and estimated arc flow vectors in the metric of D^{-1} is equal to the distance between the observed and known node flow vectors in the metric of C.

A linear combination of the arc flows may be given with a vector b as b'u. In order to estimate b'u, the linear combinations $k+h'x$ of the observed arc flows may be considered. Here k is a scalar and h is a vector which are to be chosen appropriately.

How are k and h to be chosen to provide unbiased estimators? The condition of unbiasedness is $k+h'u = b'u$ for all the vectors u that satisfy $Bu = a$. If u_0 is such a vector, then $h'(u-u_0) = b'(u-u_0)$, i.e. $(b-h)'(u-u_0) = 0$. Thus, the vector b-h has to be orthogonal to N(B). According to (12) there exists a vector g such that $b-h = B'g$. Hence $k = (b-h)'u = g'a$, and it follows that the unbiased linear combinations can be written

$$k + h'x = b'x - g'(Bx - a) \tag{38}$$

for arbitrary vectors g. Especially according to (29), $g = C'BDb$ produces the

estimator $b'u^*$. This estimator will be proved to be the unique minimum variance estimator of $b'u$ in the class of unbiased linear combinations (38).

Write $b'u = b'u_0 + (H^{-1}b)'v$ and put $c = H^{-1}b$. Consider the problem of estimating $c'v$ with any linear combination $d'y$. The condition of unbiasedness $E\, d'y = c'v$ can be formulated as $c-d \perp N(BH^{-1})$ or, according to (12), $c-d = H^{-1}B'g$ for some vector g. Expressed in the original variables, this means that $b'u$ is estimated by $b'u_0 + d'y = b'u_0 + d'H(x-u_0)$. The unbiased estimators are obtained with $d = c - H^{-1}B'g$. Substitution and simplification give the estimator $b'x - g'(Bx-a)$, i.e. the linear combinations as given by (38).

Moreover, the variance becomes $\mathrm{Var}\, d'y = \| d \|^2 = \| Pd \|^2 + \| d-Pd \|^2 \geq \geq \| Pd \|^2$ for symmetric and idempotent matrices P, with equality if, and only if, $d = Pd$. Choose P according to (34) corresponding to the projection upon $N(BH^{-1})$. The $P(c-d) = 0$, i.e. $Pd = Pc$, and it follows that $\mathrm{Var}\, d'y \geq \geq \| Pc \|^2 = c'Pc$, with equality if, and only if, $d = Pc$. The unique optimal estimator is $d'y = c'Py = c'v^*$. Expressed in the original variables, this means that for every vector g

$$\mathrm{Var}\, [b'x - g'(bx-a)] \geq b'(D-DB'CBD)b \quad , \tag{39}$$

with equality for, and only for, the estimator $b'u^*$.

4. FLOW ESTIMATION IN COMPLETE GRAPHS

4.1 Complete graphs

Consider a complete graph with n nodes. There are measurements of the net arc flows pertaining to all the $\binom{n}{2}$ pairs of nodes. The node flows are known for m out of the n nodes. A graph of this kind with $n=6$ and $m=4$ is shown in Figure 7. If the arcs between the black nodes are omitted, there remains a graph with $r = \binom{m}{2} + m(n-m) = \binom{n}{2} - \binom{n-m}{2}$ arcs. Each black node may then be replaced by m terminal nodes. The modified version of the graph in Figure 7 is shown in Figure 8.

A generalization of this kind of graph consists of the connected graphs with complete arc connections between the white nodes and no arc connections

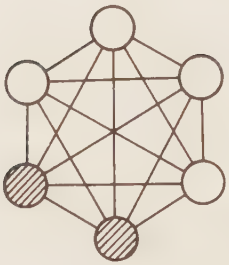


Figure 7

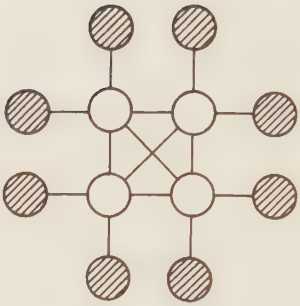


Figure 8

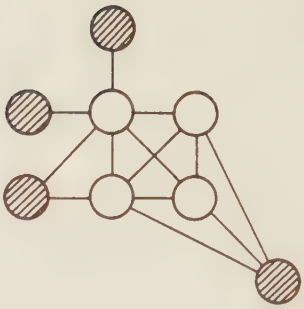


Figure 9

between the black nodes. An example is shown in Figure 9. If such a graph has m white nodes with the local degrees $r_1, r_2, \ldots, r_m$, then the number n of nodes and the number r of arcs satisfy the following relations:

$$r = \binom{m}{2} + \sum_{i=1}^{m} (r_i - m + 1) = \sum_{i=1}^{m} r_i - \binom{m}{2} \quad , \tag{40}$$

$$m - 1 \leq r_i \leq n-1 \text{ for } i = 1, 2, \ldots, m \ , \tag{41}$$

$$1 + \max_{1 \leq i \leq m} r_i \leq n \leq r - m(m-3)/2 \ . \tag{42}$$

The last equality holds if, and only if, all the black nodes are terminal nodes. The statements easily follow, if it is observed that the number of black nodes is at least $\max_i (r_i - m + 1)$ and at most $\sum_i (r_i - m + 1)$ depending upon the number of black nodes that are terminal nodes.

4.2 Explicit solutions

Explicit solutions will be derived for the flow estimation problem of the connected graphs that have complete arc connections between the white nodes and no arc connections between the black nodes. All the measurements are supposed to have the same variance, say $\sigma_{ij}^2 = 1$ for $(i,j) \in M$.

According to the earlier results, the flow estimation problem is explicitly

solved if the equation system (32) can be explicitly solved. In this case, the equation system (32) becomes

$$z_i = r_i\lambda_i - \sum_{j\neq i} \lambda_j \tag{43}$$

for $i = 1,2,\ldots,m$. By arrangement,

$$z_i = (1+r_i)\lambda_i - \lambda(\omega) \tag{44}$$

for $i = 1,2,\ldots,m$. If (44) is divided by $1+r_i$, and if summation is then carried out, it follows that

$$\sum_{i=1}^{m} z_i/(1+r_i) = [1 - \sum_{i=1}^{m} 1/(1+r_i)]\lambda(\omega) . \tag{45}$$

If $\max r_i > m-1$, i.e. if there is a black node, (45) may be solved for $\lambda(\omega)$. Substitution in (44) provides the solution

$$\lambda_i = \frac{1}{1+r_i}\left[z_i + \frac{\sum z_j/(1+r_j)}{1 - \sum 1/(1+r_j)}\right] \tag{46}$$

for $i = 1,2,\ldots,m$. If $\max r_i = m-1$, i.e. if there is no black node present, $\lambda(\omega)$ and λ_i can not be uniquely determined. However, in that case it is sufficient to know $\lambda_j - \lambda_i$ for $(i,j) \in M$. Since $r_i = m-1$ for $i = 1,2,\ldots,m$, it follows from (44) that

$$\lambda_j - \lambda_i = (z_j - z_i)/m \tag{47}$$

for $(i,j) \in M$.

In the particular case with $r_i = k+m-1$ for $i = 1,2,\ldots,m$, i.e. the same number $k > 0$ of black nodes at each white node, the formula (46) simplifies to

$$\lambda_i = z_i/(k+m) + z(\omega)/k(k+m) \tag{48}$$

for $i = 1, 2, \ldots, m$.

4.3 Application to a problem of class transitions

Suppose that each one of N individuals belongs to exactly one of n different classes. The classes may be states, newspapers, political parties, etc. The individuals may change from one class to another, and N_{ij} denotes the number of individuals that make a transition from the class i to the class j during a certain interval of time.

There is complete knowledge about the changes of the total class sizes during the interval of time, i.e.

$$\sum_{j=1}^{n} (N_{ij} - N_{ji}) = N_{i\cdot} - N_{\cdot i} \tag{49}$$

are known for $i = 1, 2, \ldots, n$. But the distribution of the changes in the different classes is unknown. For $i \neq j$, the transition numbers N_{ij} are estimated by independent estimators $\hat{N}_{ij}$. These estimators are assumed to be without systematic errors and to have the same variance, which for simplicity is put equal to 1. The more general and realistic cases may be dealt with by the same method. The provided estimators $\hat{N}_{ij}$ are usually not consistent with the total class changes. Consistent estimators N_{ij}^{*} of N_{ij} are given according to (33) and (47) as

$$N_{ij}^{*} = \hat{N}_{ij} + [(\hat{N}_{j\cdot} - \hat{N}_{\cdot j}) - (\hat{N}_{i\cdot} - \hat{N}_{\cdot i}) - (N_{j\cdot} - N_{\cdot j}) + (N_{i\cdot} - N_{\cdot i})]/n \tag{50}$$

for $i \neq j$, $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, n$. Here

$$\hat{N}_{i\cdot} = \sum_{j \neq i} \hat{N}_{ij} \text{ and } \hat{N}_{\cdot i} = \sum_{j \neq i} \hat{N}_{ji} \tag{51}$$

for $i = 1, 2, \ldots, n$.

5. FLOW ESTIMATION IN COMPLETE BIPARTITE GRAPHS

5.1 Bipartite graphs

A bipartite graph has a set of nodes Ω that consists of two disjoint subsets Ω_1 and Ω_2. No arc connects two nodes in the same subset. The graph has $n = n_1 + n_2$ nodes of which n_1 belong to Ω_1 and n_2 to Ω_2. The number of arcs is at most $n_1 n_2$. The white nodes constitute a subset $\omega \subset \Omega$ with $m = m_1 + m_2$ nodes of which m_1 belong to Ω_1 and m_2 to Ω_2. Put $\omega_1 = \omega \cap \Omega_1$ and $\omega_2 = \omega \cap \Omega_2$. In Figure 10 there is an example of a bipartite graph with $n = 16$ and $m = 11$. The same graph has been drawn in Figure 11 with the two parts Ω_1 and Ω_2 separated.

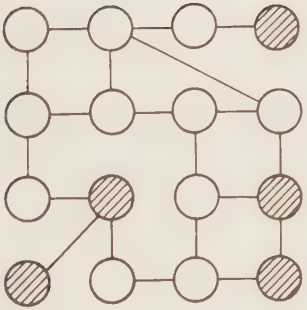


Figure 10

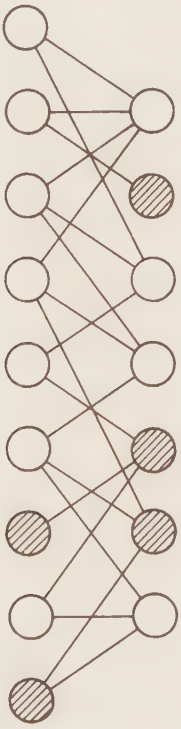


Figure 11

5.2 Complete bipartite graphs

If every node in Ω_1 has an arc connection with every node in Ω_2 the graph has $n_1 n_2$ arcs and is called a complete bipartite graph. The modified graph that arises when the arcs between the black nodes are omitted is a bipartite graph. The modified graph has complete arc connections between the white nodes in Ω_1 and Ω_2. Every white node in $\Omega_1 (\Omega_2)$ has arc connections with all the black nodes in $\Omega_2 (\Omega_1)$.

If it is imagined that the generated and absorbed flows are measured (instead of known or unknown) at some nodes, a generalization of the complete bipartite graphs becomes of interest. The generalized graph is again a bipartite graph. An example is shown in Figure 12. The generalized graph can be characterized in the following way. Every white node in Ω_1 has arc connections with every white node in Ω_2. A certain black node in $\Omega_1 (\Omega_2)$ has arc connections with some of the white nodes in $\Omega_2 (\Omega_1)$. Each one of the other black nodes in $\Omega_1 (\Omega_2)$ has arc connections with every white node in $\Omega_2 (\Omega_1)$.

A further generalization leads to arbitrary bipartite graphs where the subgraph of the white nodes is a complete bipartite graph and where there are no arcs connecting black nodes. Consider such a graph in which the white nodes have the local degrees r_i for $i \in \omega$. The number n of nodes and the number r of arcs satisfy the relations:

$$r = \sum_{i \in \omega_1} r_i + \sum_{j \in \omega_2} r_j - m_1 m_2 , \tag{52}$$

$$m_2 \leq r_i \leq n_2 \text{ for } i \in \omega_1 , \tag{53}$$

$$m_1 \leq r_j \leq n_1 \text{ for } j \in \omega_2 , \tag{54}$$

$$\max_{i \in \omega_1} r_i + \max_{j \in \omega_2} r_j \leq n \leq m_1 + m_2 - m_1 m_2 + r . \tag{55}$$

The last equality applies if, and only if, each black node is a terminal node.

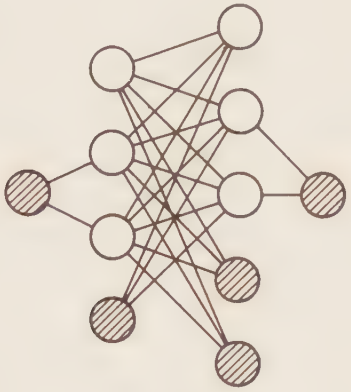


Figure 12

5.3 Explicit solutions

Explicit solutions will be found for the bipartite graphs where the subgraph of the white nodes is a complete bipartite graph, and where there are no arcs connecting black nodes. All the measurements are assumed to have the same variance, say $\sigma_{ij}^2 = 1$ for $(i,j) \in M$.

The equation system (32) of the flow estimation problem becomes

$$\left.\begin{aligned} z_i &= r_i\lambda_i - \sum_{j\in\omega_2} \lambda_j \text{ for } i \in \omega_1 \ , \\ z_j &= r_j\lambda_j - \sum_{i\in\omega_1} \lambda_i \text{ for } j \in \omega_2 \ . \end{aligned}\right\} \tag{56}$$

Use the abbreviations $\bar{z}_i = z_i/r_i$ and $K_i = 1/r_i$ for $i \in \omega$, and use the notation convention introduced earlier:

$$\lambda(S) = \sum_{i\in S} \lambda_i \ , \quad \bar{z}(S) = \sum_{i\in S} \bar{z}_i \ , \ K(S) = \sum_{i\in S} K_i \ . \tag{57}$$

It follows from (56) that

$$\left.\begin{aligned} \bar{z}(\omega_1) &= \lambda(\omega_1) - K(\omega_1)\,\lambda(\omega_2) \ , \\ \bar{z}(\omega_2) &= \lambda(\omega_2) - K(\omega_2)\,\lambda(\omega_1) \ . \end{aligned}\right\} \tag{58}$$

The equation system (58) has a solution $\lambda(\omega_1)$ and $\lambda(\omega_2)$ if, and only if, $1 - K(\omega_1)\,K(\omega_2) \neq 0$. This condition can be written

$$\sum_{i\in\omega_1}\sum_{j\in\omega_2} 1/r_i r_j \neq 1. \tag{59}$$

From the inequalities (53) and (54) it follows that (59) applies if, and only if, the graph has at least one black node. This also follows from the general results of Section 3 above.

When there is at least one black node, it follows from (58) that

$$\left.\begin{aligned}\lambda(\omega_1) &= [\bar{z}(\omega_1) + K(\omega_1)\,\bar{z}(\omega_2)]/[1 - K(\omega_1)\,K(\omega_2)]\,,\\ \lambda(\omega_2) &= [\bar{z}(\omega_2) + K(\omega_2)\,\bar{z}(\omega_1)]/[1 - K(\omega_1)\,K(\omega_2)]\,,\end{aligned}\right\} \tag{60}$$

and substitution in (56) provides the solution λ_i for $i \in \omega$. When there are only white nodes, there is no unique solution. In this case it suffices to get $\lambda_j - \lambda_i$ for $i \in \omega_1 = \Omega_1$ and $j \in \omega_2 = \Omega_2$. Since now $r_i = n_2$ for $i \in \Omega_1$ and $r_j = n_1$ for $j \in \Omega_2$, it follows from (56) that

$$\lambda_j - \lambda_i = z_j/n_1 - z_i/n_2 + \lambda(\Omega_1)/n_1 - \lambda(\Omega_2)/n_2 \tag{61}$$

for $i \in \Omega_1$ and $j \in \Omega_2$. From (58) it follows that

$$\lambda(\Omega_1)/n_1 - \lambda(\Omega_2)/n_2 = z(\Omega_1)/n_1 n_2 = -z(\Omega_2)/n_1 n_2\,, \tag{62}$$

and substitution in (61) gives the solution

$$\lambda_j - \lambda_i = z_j/n_1 - z_i/n_2 + z(\Omega_1)/n_1 n_2 \tag{63}$$

for $i \in \Omega_1$ and $j \in \Omega_2$. The solution (63) can be given a more elegant form by introducing the notations

$$\left.\begin{aligned} \bar{u}_{i\cdot} &= \sum_{j\in\Omega_2} u_{ij}/n_2 , & \bar{x}_{i\cdot} &= \sum_{j\in\Omega_2} x_{ij}/n_2 , \\ \bar{u}_{\cdot j} &= \sum_{i\in\Omega_1} u_{ij}/n_1 , & \bar{x}_{\cdot j} &= \sum_{i\in\Omega_1} x_{ij}/n_1 , \\ \bar{u} &= \sum_{i\in\Omega_1}\sum_{j\in\Omega_2} u_{ij}/n_1 n_2 , & \bar{x} &= \sum_{i\in\Omega_1}\sum_{j\in\Omega_2} x_{ij}/n_1 n_2 . \end{aligned}\right\} \tag{64}$$

The relationship then is that

$$\left.\begin{aligned} z_i &= n_2(\bar{x}_{i\cdot} - \bar{u}_{i\cdot}) \text{ for } i\in\Omega_1 , \\ z_j &= n_1(\bar{u}_{\cdot j} - \bar{x}_{\cdot j}) \text{ for } j\in\Omega_2 , \end{aligned}\right\} \tag{65}$$

and from (63) it follows that

$$\lambda_j - \lambda_i = \bar{u}_{i\cdot} - \bar{x}_{i\cdot} + \bar{u}_{\cdot j} - \bar{x}_{\cdot j} - \bar{u} + \bar{x} \tag{66}$$

for $i\in\Omega_1$ and $j\in\Omega_2$. Hence u_{ij} has the least squares estimator

$$u^*_{ij} = x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x} + \bar{u}_{i\cdot} + \bar{u}_{\cdot j} - \bar{u} \tag{67}$$

for $i\in\Omega_1$ and $j\in\Omega_2$.

5.4 Application to two-way contingency tables

Consider a two-way classified population with k_1 levels of the first classification factor and k_2 levels of the second classification factor. The population is divided into k_1k_2 classes, and the class indexed (i,j) consists of N_{ij} units for $i = 1,2,\ldots,k_1$ and $j = 1,2,\ldots,k_2$. The marginal frequencies associated with the first and second classification factors are denoted respectively

$$N_{i\cdot} = \sum_{j=1}^{k_2} N_{ij} \quad \text{and} \quad N_{\cdot j} = \sum_{i=1}^{k_1} N_{ij} \tag{68}$$

for $i = 1,2,\ldots,k_1$ and $j = 1,2,\ldots,k_2$. The total size of the population is denoted

$$N = \sum_{i=1}^{k_1} N_{i\cdot} = \sum_{j=1}^{k_2} N_{\cdot j} \,. \tag{69}$$

The size of the population and some or all of the sizes of the classes may be unknown. One has recourse to independent estimators of the unknown frequencies, and these are denoted $\hat{N}$, $\hat{N}_{i\cdot}$, $\hat{N}_{\cdot j}$, and $\hat{N}_{ij}$. The estimators do not usually satisfy the condition of consistency as given by (68) and (69). The problem is to get consistent estimators with good properties.

Deming and Stephan (1940) have dealt with the problem of estimating the frequencies of a contingency table with known marginal values. They start with the observed class frequencies in a random sample, i.e. a multinomial distribution on the $k_1 k_2$ classes. By applying a least squares criterion, they reduce the estimation problem to the solution of a linear equation system. Neyman (1949) and Taylor (1953) have shown that the solutions are "best asymptotically normal". In order to find easier methods of solution, Deming and Stephan (1940) and Stephan (1942) introduced and studied iterative techniques. Iterative procedures have also been studied by Brown (1959) and by Ireland and Kullback (1968). In the last paper a criterion is used that is based upon the discrimination information (Kullback, 1959). Some other papers dealing with the estimation of the frequencies in contingency tables with fixed marginal values are those of El Badry and Stephan (1955) and Thionet (1959). A geometrical description of an iterative procedure is given in a recent paper by Fienberg (1970).

The essential difference between the estimation problems of the papers mentioned and the problem dealt with here is the condition of independence. Instead of having observations following a multinomial distribution, it is assumed that there are given preliminary independent estimators which are inconsistent with the marginal restrictions. The problem is to provide consistent estimators.

The two-way contingency tables can be represented by bipartite graphs. A graph corresponding to a 2×3 table is shown in Figure 13. The graph corre-

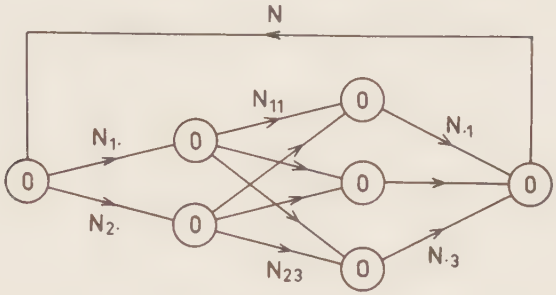


Figure 13

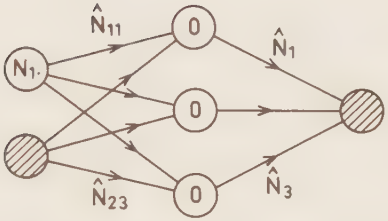


Figure 14

sponding to the case where N is unknown, $N_{1\cdot}$ is known, $N_{2\cdot}$ is unknown, N_{ij} is unknown with the estimator $\hat{N}_{ij}$, and $N_{\cdot j}$ is unknown with the estimator $\hat{N}_j$ for $i = 1,2$ and $j = 1,2,3$ is shown in Figure 14. Suppose that the variance of the estimators are known and denoted by $\mathrm{Var}\ \hat{N}_{ij} = \sigma_{ij}^2$ and $\mathrm{Var}\ \hat{N}_j = \sigma_j^2$ for $i = 1,2$ and $j = 1,2,3$. Put

$$\left.\begin{aligned} z_1 &= \hat{N}_{11} + \hat{N}_{12} + \hat{N}_{13} - N_{1\cdot} = \hat{N}_{1\cdot} - N_{1\cdot} \ , \\ z_2 &= \hat{N}_1 - \hat{N}_{11} - \hat{N}_{21} = \hat{N}_1 - \hat{N}_{\cdot 1} \ , \\ z_3 &= \hat{N}_2 - \hat{N}_{12} - \hat{N}_{22} = \hat{N}_2 - \hat{N}_{\cdot 2} \ , \\ z_4 &= \hat{N}_3 - \hat{N}_{13} - \hat{N}_{23} = \hat{N}_3 - \hat{N}_{\cdot 3} \ . \end{aligned}\right\} \qquad (70)$$

The equation system (32) of the flow estimation problem becomes

$$\left.\begin{aligned} z_1 &= \lambda_1 (\sigma_{11}^2 + \sigma_{12}^2 + \sigma_{13}^2) - \lambda_2 \sigma_{11}^2 - \lambda_3 \sigma_{12}^2 - \lambda_4 \sigma_{13}^2 \ , \\ z_2 &= \lambda_2 (\sigma_1^2 + \sigma_{11}^2 + \sigma_{21}^2) - \lambda_1 \sigma_{11}^2 \ , \\ z_3 &= \lambda_3 (\sigma_2^2 + \sigma_{12}^2 + \sigma_{22}^2) - \lambda_1 \sigma_{12}^2 \ , \\ z_4 &= \lambda_4 (\sigma_3^2 + \sigma_{13}^2 + \sigma_{23}^2) - \lambda_1 \sigma_{13}^2 \ , \end{aligned}\right\} \tag{71}$$

and the consistent estimators are given by

$$\left.\begin{aligned} N_{1j}^* &= \hat{N}_{1j} + (\lambda_{j+1} - \lambda_1) \sigma_{1j}^2 \quad \text{for } j = 1,2,3 \ , \\ N_{2j}^* &= \hat{N}_{2j} + \lambda_{j+1} \sigma_{2j}^2 \quad \text{for } j = 1,2,3 \ , \\ N_{\cdot j}^* &= \hat{N}_{\cdot j} - \lambda_{j+1} \sigma_j^2 \quad \text{for } j = 1,2,3 \ . \end{aligned}\right\} \tag{72}$$

The simple formula (66) may be applied to two-way contingency tables with known margins N, $N_{i\cdot}$, $N_{\cdot j}$ and independent estimators $\hat{N}_{ij}$ with the same variance. In this case, the consistent estimators N_{ij}^* are given as

$$N_{ij}^* = \hat{N}_{ij} - \hat{N}_{i\cdot} - \hat{N}_{\cdot j} + \hat{N} + N_{i\cdot} + N_{\cdot j} - N \ , \tag{73}$$

with

$$\hat{N} = \sum_i \hat{N}_{i\cdot} = \sum_j \hat{N}_{\cdot j} = \sum_i \sum_j \hat{N}_{ij} \ . \tag{74}$$

6. FLOW ESTIMATION IN TREES

6.1 Trees

Consider a tree graph with measurements of all the arc flows. Obviously it is no real restriction to consider only those trees which have black terminal nodes and white non-terminal nodes.

The tree graphs can be generalized to the connected graphs which are characterized by that the subgraph with the white nodes is a tree, there is no

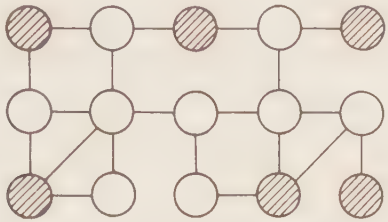


Figure 15

arc between black nodes, and there is no white terminal node. An example is provided by Figure 15. For graphs of this type, with m white nodes which have the local degrees $r_1, r_2, \ldots, r_m$, the following relations apply for the number n of nodes and the number r of arcs:

$$r = \sum_{i=1}^{m} r_i - m + 1 \quad , \tag{75}$$

$$r_i \geq 2 \text{ for } i = 1, 2, \ldots, m \quad , \tag{76}$$

$$m + b \leq n \leq r + 1 \quad . \tag{77}$$

Here b is the maximum number of black nodes connected with any white node. If, and only if, all the black nodes are terminal nodes, the graph is a tree with $n = r + 1$. For the flow estimation problem it is no restriction to assume that the graph is a tree with n nodes, of which m are white non-terminal nodes and $n - m$ are black terminal nodes.

6.2 The terminal node classes

Some concepts will be introduced that will prove to be useful later when a solution algorithm is given for the flow estimation problem for trees.

Consider an arbitrary tree with the set Ω of nodes. Let T_0 be the class of terminal nodes. If these nodes and their arcs are omitted, there remains a subtree. Let T_1 be the class of terminal nodes of the subtree. By omitting successively the terminal nodes with their arcs, the set Ω is partitioned into a

series of disjoint classes $T_0, T_1, \ldots, T_s$. The classes are called the successive terminal node classes of the tree. The last class T_s consists of either a single node or two nodes with an arc. In the first case, the tree is said to have a center, and in the second case the tree is said to have a bicenter. The class T_s is called the central class.

A tree with a center is shown in Figure 16. The nodes belonging to T_v are labeled v for $v = 0, 1, \ldots, s$. An example of a tree with a bicenter is given in Figure 17.

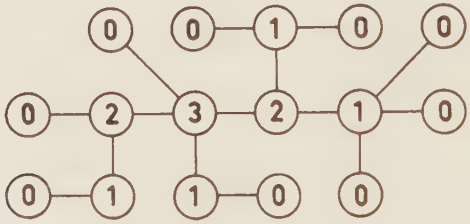


Figure 16

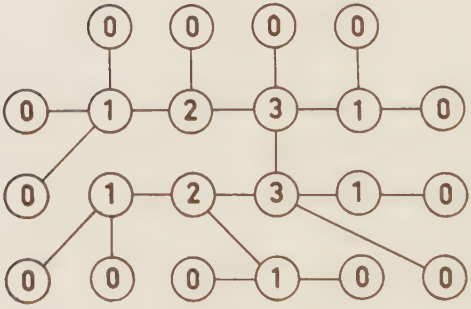


Figure 17

The following properties may be noted of the successive terminal node classes. Each node which does not belong to the central class has an arc connection with exactly one node from a more central class (i.e. a class with a higher index). There are no arcs between the nodes in the same class T_v for $v = 0, 1, \ldots, s-1$. It is therefore possible for each node $j \in \Omega - T_s$ to define g(j) as the unique node which has an arc connection with node j and which belongs to a more cen-

tral class than node j. For each node $j \in \Omega - T_0$, it is possible to define G(j) as the set of nodes which have arc connections with node j and which belong to more peripheral classes (i.e. classes with lower indices) than node j. For $v = 0, 1, \ldots, s-1$ and $j \in T_v$, it applies that

$$g(j) \in T_{v+1} \cup \ldots \cup T_s \,, \tag{78}$$

and for $v = 1, 2, \ldots, s$ and $j \in T_v$, it applies that

$$G(j) \subset T_0 \cup \ldots \cup T_{v-1} \,. \tag{79}$$

6.3 The method of solution

Consider a tree with n nodes, of which m are white non-terminal nodes and n - m are black terminal nodes. Let Ω denote the set of nodes and ω the set of white nodes. With the notation of the last subsection, $T_0 = \Omega - \omega$ and $\omega = T_1 \cup \ldots \cup T_s$. Suppose that all the measurement variances are equal, say $\sigma_{ij}^2 = 1$ for $(i,j) \in M$.

The equation system (32) of the flow estimation problem reads as follows when the tree has a center and the center is labeled 1:

$$\left.\begin{aligned} z_1 &= r_1\lambda_1 - \lambda(G(1)) \,, \\ z_j &= r_j\lambda_j - \lambda_{g(j)} - \lambda(G(j)) \text{ for } j \in \omega - T_s \,. \end{aligned}\right\} \tag{80}$$

For $j \in T_0$, $\lambda_j = 0$ as usual. When the tree has a bicenter and the central nodes are labeled 1 and 2, the equation system becomes

$$\left.\begin{aligned} z_1 &= r_1\lambda_1 - \lambda_2 - \lambda(G(1)) \,, \\ z_2 &= r_2\lambda_2 - \lambda_1 - \lambda(G(2)) \,, \\ z_j &= r_j\lambda_j - \lambda_{g(j)} - \lambda(G(j)) \text{ for } j \in \omega - T_s \,. \end{aligned}\right\} \tag{81}$$

An examination of the equations shows that λ_i for a fixed $i \in T_1$ is given as a linear function of $\lambda_{g(i)}$ with $g(i) \in T_2$. Moreover, λ_j for a fixed $j \in T_2$ is given as a linear combination of $\lambda_{g(j)}$ with $g(j) \in T_3$ and the λ_i with $i \in G(j) \subset T_0 \cup T_1$. Since $i \in G(j)$ is equivalent to $j = g(i)$ all these λ_i are expressible as linear functions of λ_j. It follows that λ_j for each $j \in T_2$ can be given as a linear function of $\lambda_{g(j)}$. By continuing in this way it appears that the λ_j for $j \in T_{s-1}$ are expressible as linear functions of $\lambda_{g(j)}$ with $g(j) \in T_s$. Substitution in the λ_1 equation of (80) or in the λ_1 and λ_2 equations of (81) gives the central solution λ_1 or λ_1, λ_2, respectively. Then, by using the linear relationships, the λ_j for $j \in T_{s-1}, T_{s-2}, \ldots, T_1$ are successively obtained.

The equation systems (80) or (81) can be solved by a solution algorithm based upon the present reasoning. In the next subsection it will be proved that a linear relationship $\lambda_j = A_j + B_j \lambda_{g(j)}$ holds for $j \in \omega - T_s$. Computational algorithms will be given for the coefficients A_j and B_j. In the following subsections the central solution λ_j for $j \in T_s$ will be obtained, and an algorithm will be given that provides the solution λ_j for $j \in \omega - T_s$.

6.4 The induction step

Define $A_i = B_i = 0$ for $i \in T_0$. For $i \in T_1$ it follows from (80) or (81) that $\lambda_i = A_i + B_i \lambda_{g(i)}$, with $A_i = z_i/r_i$ and $B_i = 1/r_i$. For $j \in T_2$,

$$z_j = r_j \lambda_j - \lambda_{g(j)} - \sum_{i \in G(j)} (A_i + B_i \lambda_j) = r_j \lambda_j - \lambda_{g(j)} - A(G(j)) - B(G(j))\lambda_j . \tag{82}$$

Hence

$$\lambda_j = A_j + B_j \lambda_{g(j)} \tag{83}$$

with

$$A_j = [z_j + A(G(j))]/[r_j - B(G(j))] , \tag{84}$$

$$B_j = 1/[r_j - B(G(j))] . \tag{85}$$

The denominators are positive since G(j) has at most r_j-1 nodes belonging to $T_0 \cup T_1$, and these nodes have $B_i<1$. It follows that $B(G(j)) < r_j - 1$, which in turn implies that $B_j < 1$. The relationships (83), (84), and (85) are valid for $j \in T_2$. By inspection it follows that they are also valid for $j \in T_1$, i.e. they are valid for $j \in T_1 \cup T_2$.

Suppose that (83), (84), and (85) are correct for $j \in T_1 \cup \ldots \cup T_v$, and consider $j \in T_{v+1}$. Here $v \leq s-2$. From (80) or (81) it then follows that the equation (82) applies. Thus, the results in (83), (84), and (85) are proved for $j \in T_{v+1}$, i.e. for $j \in T_1 \cup \ldots \cup T_{v+1}$. By induction it follows that (83), (84), and (85) are correct for $j \in T_1 \cup \ldots \cup T_{s-1}$.

6.5 The central solution

If the central class consists of node 1 only, it follows from (80) and (83) with $j \in T_{s-1}$ that

$$z_1 = r_1\lambda_1 - \sum_{j \in G(1)} (A_j + B_j\lambda_1) \, . \tag{86}$$

Hence

$$\lambda_1 = [z_1 + A(G(1))]/[r_1 - B(G(1))] = A_1 \tag{87}$$

in accordance with (84).

If the central class consists of the two nodes 1 and 2, it follows from (81) and (83) with $j \in T_{s-1}$ that

$$\left.\begin{aligned} z_1 &= r_1\lambda_1 - \lambda_2 - \sum_{j \in G(1)} (A_j + B_j\lambda_1) \, , \\ z_2 &= r_2\lambda_2 - \lambda_1 - \sum_{j \in G(2)} (A_j + B_j\lambda_2) \, . \end{aligned}\right\} \tag{88}$$

The equation system (88) can be written as

$$\left.\begin{aligned} \lambda_1 &= A_1 + B_1\lambda_2 \ , \\ \lambda_2 &= A_2 + B_2\lambda_1 \ , \end{aligned}\right\} \tag{89}$$

with A_1, A_2, B_1, B_2 given according to (84) and (85). According to (85) it follows by induction that $B_1 < 1$ and $B_2 < 1$. Thus, the equation system (89) has the unique solution

$$\left.\begin{aligned} \lambda_1 &= (A_1 + B_1A_2)/(1 - B_1B_2) \ , \\ \lambda_2 &= (A_2 + B_2A_1)/(1 - B_1B_2) \ . \end{aligned}\right\} \tag{90}$$

6.6 The algorithm

The results above imply that the equation systems (80) or (81) can be solved in the following way.

Put $B_i = 0$ for $i \in T_0$, and compute B_j successively for $j \in T_1, T_2, \ldots, T_s$ according to (85).

Put $A_i = 0$ for $i \in T_0$, and compute A_j successively for $j \in T_1, T_2, \ldots, T_s$ according to (84).

If T_s consists of only one node, the central solution is $\lambda_1 = A_1$. If T_s consists of two nodes, the central solutions are given by (90).

Compute λ_j successively for $j \in T_{s-1}, T_{s-2}, \ldots, T_1$ according to (83).

6.7 An example

The algorithm will be demonstrated by a simple example. The tree in Figure 18 has its nodes labeled from the central class towards the terminal

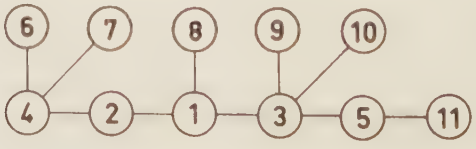


Figure 18

nodes, i.e. the nodes of T_v have lower numbers than the nodes of T_{v-1} for $v = 1, 2, \ldots, s$. The tree has only one center. The basic data are shown in the following table.

j	g(j)	$\omega \cap G(j)$	r_j
1		2, 3	3
2	1	4	2
3	1	5	4
4	2		3
5	3		2

The equation system (80) becomes

$$\left.\begin{aligned} z_1 &= 3\lambda_1 - \lambda_2 - \lambda_3 \ , \\ z_2 &= 2\lambda_2 - \lambda_1 - \lambda_4 \ , \\ z_3 &= 4\lambda_3 - \lambda_1 - \lambda_5 \ , \\ z_4 &= 3\lambda_4 - \lambda_2 \ , \\ z_5 &= 2\lambda_5 - \lambda_3 \ . \end{aligned}\right\} \tag{91}$$

The coefficients are successively calculated according to the algoritm as

$$\left.\begin{aligned} B_5 &= 1/2 \ , \\ B_4 &= 1/3 \ , \\ B_3 &= 1/(4-B_5) = 2/7 \ , \\ B_2 &= 1/(2-B_4) = 3/5 \ , \\ B_1 &= 1/(3-B_2-B_3) = 35/74 \ , \end{aligned}\right\} \tag{92}$$

$$
\left.
\begin{aligned}
A_5 &= z_5/2 \ , \\
A_4 &= z_4/3 \ , \\
A_3 &= B_3(z_3+A_5) = (2z_3+z_5)/7 \ , \\
A_2 &= B_2(z_2+A_4) = (3z_2+z_4)/5 \ , \\
A_1 &= B_1(z_1+A_2+A_3) = (35z_1+21z_2+10z_3+7z_4+5z_5)/74 \ ,
\end{aligned}
\right\} \qquad (93)
$$

and the solution becomes

$$
\left.
\begin{aligned}
\lambda_1 &= A_1 = (35z_1+21z_2+10z_3+7z_4+5z_5)/74 \ , \\
\lambda_2 &= A_2 + B_2\lambda_1 = (21z_1+57z_2+6z_3+19z_4+3z_5)/74 \ , \\
\lambda_3 &= A_3 + B_3\lambda_1 = (10z_1+6z_2+24z_3+2z_4+12z_5)/74 \ , \\
\lambda_4 &= A_4 + B_4\lambda_2 = (7z_1+19z_2+2z_3+31z_4+z_5)/74 \ , \\
\lambda_5 &= A_5 + B_5\lambda_3 = (5z_1+3z_2+12z_3+z_4+43z_5)/74 \ .
\end{aligned}
\right\} \qquad (94)
$$

6.8 Generalization

The case of unequal measurement variances can be dealt with by an algorithm that works in complete analogy with the one described above. The reasoning and the proofs are the same and may be omitted. Use the abbreviation $\sigma^2_{jk} = \sigma^2_{kj} = \sigma^2_j$ if $k = g(j)$ and $j \notin T_s$. When there are two centers, put $\sigma^2_{21} = \sigma^2_{12} = \sigma^2_2 = \sigma^2_1$. In the general case, the equation system (80) is replaced by

$$
\left.
\begin{aligned}
z_1 &= \sigma^2(G(1))\lambda_1 - \sum_{i\in G(1)} \sigma^2_i \lambda_i \ , \\
z_j &= [\sigma^2_j + \sigma^2(G(j))]\,\lambda_j - \sigma^2_j \lambda_{g(j)} - \sum_{i\in G(j)} \sigma^2_i \lambda_i \quad \text{for } j \in \omega - T_s \ ,
\end{aligned}
\right\} \qquad (95)
$$

and the equation system (81) is replaced by

$$\left.\begin{aligned} z_1 &= [\sigma_1^2 + \sigma^2(G(1))]\,\lambda_1 - \sigma_1^2\,\lambda_2 - \sum_{i\in G(1)} \sigma_i^2\,\lambda_i \ , \\ z_2 &= [\sigma_2^2 + \sigma^2(G(2))]\,\lambda_2 - \sigma_2^2\,\lambda_1 - \sum_{i\in G(2)} \sigma_i^2\,\lambda_i \ , \\ z_j &= [\sigma_j^2 + \sigma^2(G(j))]\,\lambda_j - \sigma_j^2\,\lambda_{g(j)} - \sum_{i\in G(j)} \sigma_i^2\lambda_i \ \text{ for } j\in\omega\text{-}T_s \ . \end{aligned}\right\} \quad (96)$$

The relations (84) and (85), which define the coefficients of the algorithm, are changed to

$$A_j = (z_j + \sum_{i\in G(j)} \sigma_i^2\,A_i)/[\sigma_j^2 + \sigma^2(G(j)) - \sum_{i\in G(j)} \sigma_i^2\,B_i] \ , \quad (97)$$

$$B_j = \sigma_j^2/[\sigma_j^2 + \sigma^2(G(j)) - \sum_{i\in G(j)} \sigma_i^2\,B_i] \quad (98)$$

for $j\in\omega$. The central solution or the central solutions are obtained by application of (87) or (90) respectively. Then the linear relationships (83) are applied to provide the complete solution.

6.9 Notes

It may be noted that the coefficients B_j are not stochastic variables but depend only upon the tree structure and the variances of the measurements. The coefficients A_j are stochastic variables and depend upon both the flows and the tree structure.

The introduction of the successive terminal node classes implies an orientation of the arcs of the tree from the central class T_s towards the terminal node class T_0. The orientation is given by the adjacency operator G. It may be said that the orientation is achieved by selecting a center as the root of the tree and by directing the arcs away from the root. However, it is not essential for the algorithm that the root is a center. An arbitrary node may be selected as the root of the tree, and the adjacency operator G is then defined by use of the orientation from that root. In a certain sense it is best to take a center as the

root. The center or the centers are situated in the middle of the longest arc chain of the tree. For a proof of this property of the centers, reference is made to König (1936). It follows that the least possible number of iterations in the algorithm is obtained when the root is a center.

6.10 Regular trees

The flow estimation problem can be explicitly solved for a certain kind of trees which will be called regular. This case is taken up, since it is possible to give the solution an interesting interpretation that may be of value in other cases.

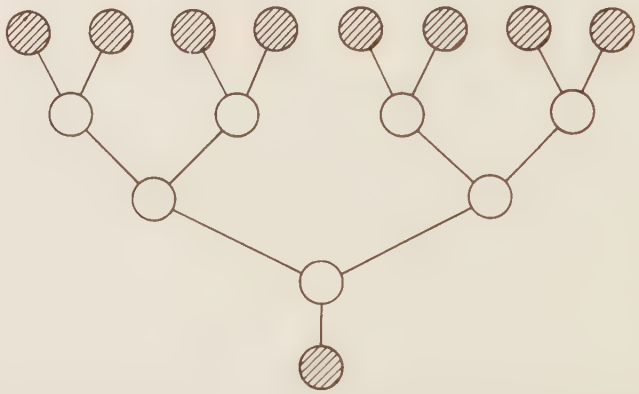


Figure 19

A tree will be called regular if it has the following structure. One node is arc connected with one terminal node and t other nodes. Each one of these t nodes is arc connected with t other nodes, etc. Each one of the nonterminal nodes consequently has $t+1$ arcs. An s-stage tree has in total $r = 1 + t + t^2 + \ldots + t^s$ arcs and $r+1$ nodes, of which $1+t^s$ are terminal nodes. A regular tree with $s=3$ and $t=2$ is shown in Figure 19.

Assume that $t > 1$. Then T_0 consists of the $1+t^s$ terminal nodes, T_1 consists of t^{s-1} nodes, T_2 consists of t^{s-2} nodes, ..., T_s consists of one node. When the variances of the measurements are all equal to one, we have by application of the algorithm

$$B_j = 1 - t^v/(1 + t + \ldots + t^v) \quad , \tag{99}$$

and

$$A_j = [(1 + t + \ldots + t^{v-1})\, z_j + (1 + t + \ldots + t^{v-2})\, z(G(j)) + \ldots + \\ + (1 + t)\, z(G_{v-2}(j)) + z(G_{v-1}(j))]/(1 + t + \ldots + t^v) \tag{100}$$

for $j \in T_v$ and $v = 1, 2, \ldots, s$. For the center, node 1, $G(1) = T_{s-1}$, $G_2(1) = T_{s-2}$, $\ldots$, $G_{s-1}(1) = T_1$, and it follows that

$$\lambda_1 = A_1 = \sum_{v=1}^{s} (1 + t + \ldots + t^{v-1})\, z(T_v)/(1 + t + \ldots + t^s) = \\ = \sum_{v=1}^{s} t^{v-1} z(T_v \cup \ldots \cup T_s)/(1 + t + \ldots + t^s) \, . \tag{101}$$

If the terminal node at node 1 is labeled node 0, the net arc flow from node 0 to node 1, the so called root flow, is estimated by

$$u_{01}^* = x_{01} + \lambda_1 \, . \tag{102}$$

In order to interpret the estimator (102), a transcription will be needed. Put $T_{s+1} = T_0$ and $x(S, T) = \sum_{i \in S} \sum_{j \in T} x_{ij}$. Then

$$z(T_v) = x(T_v, T_{v-1}) - x(T_{v+1}, T_v) - a(T_v) \tag{103}$$

for $v = 1, 2, \ldots, s$. Especially, $x(T_0, T_s) = x_{01}$. From (103) it follows that

$$z(T_v \cup \ldots \cup T_s) = x(T_v, T_{v-1}) - x_{01} - a(T_v \cup \ldots \cup T_s) \tag{104}$$

for $v = 1, 2, \ldots, s$. Put

$$y_{s-v+1} = x(T_v, T_{v-1}) - a(T_v \cup \ldots \cup T_s) \tag{105}$$

for $v = 1,2,\ldots,s$. The variable y_1 is the observed net flow from T_s to T_{s-1} minus the known net flow from T_s. The variable y_2 is the observed net flow from T_{s-1} to T_{s-2} minus the known net flow from $T_{s-1} \cup T_s$, etc. By using the fact that $a(T_v \cup \ldots \cup T_s) = u(T_v, T_{v-1}) - u_{01}$ and $E\, x(T_v, T_{v-1}) = u(T_v, T_{v-1})$ it follows that $E\, y_v = u_{01}$ for $v = 1,2,\ldots,s$. As there are t^{s-v+1} arcs between T_v and T_{v-1}, it follows that $\mathrm{Var}\, y_v = t^v$ for $v = 1,2,\ldots,s$. Substitution of (104) and (105) in (101) and (102) gives after simplification

$$u_{01}^* = (x_{01} + y_1/t + y_2/t^2 + \ldots + y_s/t^s)/(1 + 1/t + 1/t^2 + \ldots + 1/t^s) \ , \qquad (106)$$

i.e. u_{01}^* is equal to the unbiased linear combination of $x_{01}, y_1, y_2, \ldots, y_s$ that has the least possible variance.

6.11 A non-optimal estimator

A study of the estimators that are obtained in analogy with (106) for more general trees would seem to be indicated.

Consider a tree that has one center, node 1, which has a terminal node, node 0, at the distance of one arc and which has all the other-terminal nodes of the tree at the distance of s arcs from the center. Such a tree is exemplified in Figure 20.

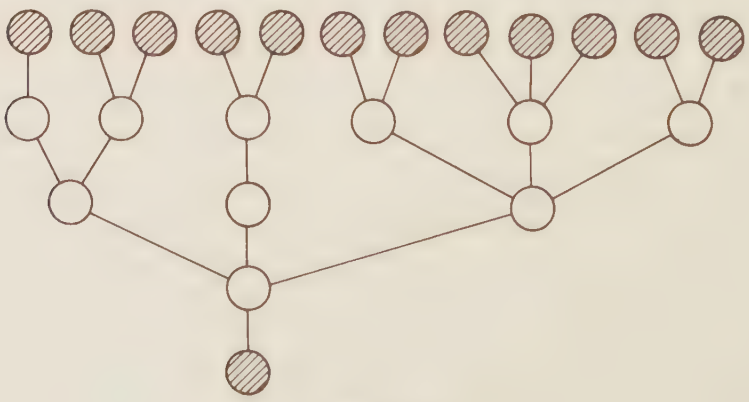


Figure 20

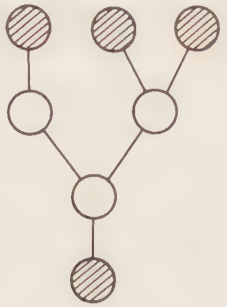


Figure 21

For a tree of this kind, the variables $y_0 = x_{01}$ and $y_1, y_2, \ldots, y_s$ defined according to (105) are independent stochastic variables with $E\, y_v = u_{01}$ for $v = 0, 1, \ldots, s$. Put $\mathrm{Var}\, y_v = V_v$ for $v = 0, 1, \ldots, s$. An unbiased estimator of u_{01} is provided by the linear minimum variance combination of these variables, i.e. by

$$u_{01}^* = \left(\sum_{v=0}^{s} y_v / V_v \right) / \left(\sum_{v=0}^{s} 1/V_v \right) . \qquad (107)$$

In general, the estimator (107) does not coincide with the least squares estimator. This is easily found by examination of some appropriate example. Consider for instance the tree of Figure 21 with unit variances of all the measurements. The position is that u_{01} has a least squares estimator that has the variance 7/13, but the estimator according to (107) has the variance 6/11.

6.12 Application to nested classifications

Consider a population with N units that is composed of k disjoint subpopulations consisting of $N_1, N_2, \ldots N_k$ units. The i-th subpopulation is in turn composed of k_i disjoint subsubpopulations consisting of N_{ij} units for $j = 1, 2, \ldots, k_i$ and $i = 1, 2, \ldots, k$. Consequently, $\sum_j N_{ij} = N_i$ for $i = 1, 2, \ldots, k$, and $\sum_i N_i = N$.

The nested classification described may be represented by a tree graph. Suppose that there are independent unbiased estimators $\hat{N}$, $\hat{N}_i$, and $\hat{N}_{ij}$ of N, N_i, and N_{ij}, respectively, for $j = 1, 2, \ldots, k_i$ and $i = 1, 2, \ldots, k$. In Figure 22 such

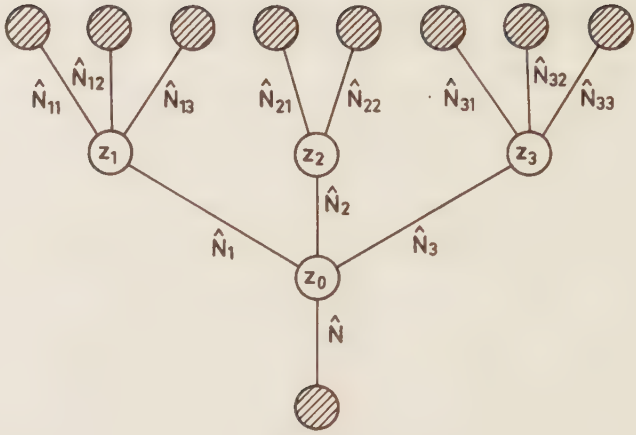


Figure 22

a graph is shown together with the observed node and arc flows. The observed node flows are defined from

$$\left.\begin{aligned} z_0 &= \sum_i \hat{N}_i - \hat{N} \ , \\ z_i &= \sum_j \hat{N}_{ij} - \hat{N}_i \quad \text{for } i = 1,2,\ldots,k \ . \end{aligned}\right\} \tag{108}$$

Suppose that $\hat{N}$, $\hat{N}_i$, and $\hat{N}_{ij}$ have the variances σ^2, σ_i^2, and σ_{ij}^2, respectively, for $j = 1,2,\ldots,k_i$ and $i = 1,2,\ldots,k$. Then the least squares estimators are given by

$$\left.\begin{aligned} N^* &= \hat{N} + \sigma^2 \lambda_0 \ , \\ N_i^* &= \hat{N}_i + \sigma_i^2 (\lambda_i - \lambda_0) \ , \\ N_{ij}^* &= \hat{N}_{ij} - \sigma_{ij}^2 \lambda_i \ , \end{aligned}\right\} \tag{109}$$

for $j = 1,2,\ldots,k_i$ and $i = 1,2,\ldots,k$. The unknown $\lambda_0, \lambda_1, \ldots, \lambda_k$ are easily obtained from the following equation system:

$$\left.\begin{aligned} z_0 &= (\sigma^2 + \sum_i \sigma_i^2)\lambda_0 - \sum_i \sigma_i^2 \lambda_i \ , \\ z_i &= (\sigma_i^2 + \sum_j \sigma_{ij}^2)\lambda_i - \sigma_i^2 \lambda_0 \quad \text{for } i = 1,2,\ldots,k. \end{aligned}\right\} \tag{110}$$

Chapter 7

INFERENCE FROM STOCHASTIC CONTACT GRAPHS

SYNOPSIS

1. Stochastic contact graphs. A simple type of graph with stochastically occurring arcs is defined, and two different interpretations are provided. The stochastic graph may be considered as a sample graph obtained by arc sampling from an unknown population graph. The stochastic graph may also be considered as the result of a random choice of contacts from a given set of possible contacts.

2. Undirected complete population graphs. The population graph or the graph consisting of the possible contacts is assumed to be undirected and complete. The properties of the stochastic contact graph are analyzed. Particular attention is paid to the isolate problem and the clique problem of sociometry. Two problems of stochastic connectedness are studied which have applications in reliability theory.

3. Directed complete population graphs. The properties of a stochastic contact graph are analyzed when the population graph is a directed complete graph. The reciprocal choice problem, the isolate problem, and the clique problem of

sociometry are considered. The most common versions of these problems are included in the present context by introducing certain conditional probability distributions. Many generalizations and simple proofs are provided of various results from sociometry.

4. General population graphs. Stochastic contact graphs are studied when the population graph is an arbitrary undirected or directed graph. The moment properties of the distribution of the local degrees in the sample graph are investigated, and the corresponding distribution in the population graph is estimated. The isolate problem is treated. An application to quality control of grammophone records is discussed.

1. STOCHASTIC CONTACT GRAPHS

1.1 Arc sampling

The notations of a population graph that have been introduced in Chapter 4 will also be used in the present chapter. Thus, consider an arbitrary graph with the node set $\Omega = \{1,2,\ldots,N\}$ and the arc sets M_{ij} for $i \in \Omega$ and $j \in \Omega$. The total arc set is denoted M, and it has R arcs. The arc indicator matrix is denoted A, the arc frequency matrix is denoted C, and the vectors of the local out-degrees and in-degrees are denoted a and b respectively.

From the arc set M a subset $T \subset M$ is chosen by random sampling. The sample graph has the node set Ω and the arc sets $T \cap M_{ij}$ for $i \in \Omega$ and $j \in \Omega$. Consequently, the sample graph is a stochastic partial graph of the population graph. The arc indicator matrix of the sample graph is denoted X, the arc frequency matrix is denoted Z, and the vectors of the local out-degrees and in-degrees are denoted x and y respectively.

Consider a sampling procedure according to which each arc has probability p of being chosen and different arcs are selected independently, i.e. a Bernoulli sampling of arcs with selection probability p. The sample graphs obtained in this way will be called stochastic contact graphs. Put $q = 1-p$.

In the directed case, the arc frequencies Z_{ij} are independent and binomial (C_{ij},p) for $i \in \Omega$ and $j \in \Omega$. Binomial $(0,p)$ has to be interpreted as the constant zero. In the undirected case, $Z_{ij} = Z_{ji}$ are independent and binomial (C_{ij},p) for $i \leq j$. Moreover, the arc indicators X_{ij} are independent and Bernoulli $(1-q^{C_{ij}})$ for all i and j in the directed case and for $i \leq j$ in the undirected case. Particularly, when the population graph has no multiple arcs it follows that $X_{ij} = Z_{ij}$ are independent and binomial (A_{ij},p) for all i and j in the directed case and for $i \leq j$ in the undirected case.

The total arc frequency r of the stochastic contact graph is binomial (R,p). If the sampling procedure is conditioned by a fixed value r of the total arc frequency in the sample graph, the conditional sampling procedure will become identical with random sampling of r arcs without replacement from the set M of arcs in the population graph. In the directed case, the various outcomes Z with

$\sum_i \sum_j Z_{ij} = r$ will have the conditional probability

$$\prod_i \prod_j \binom{C_{ij}}{Z_{ij}} / \binom{R}{r} . \tag{1}$$

In the undirected case, the various outcomes Z with $\sum_{i \le j} \sum Z_{ij} = r$ will have the conditional probability

$$\prod_{i \le j} \prod \binom{C_{ij}}{Z_{ij}} / \binom{R}{r} . \tag{2}$$

The local out-degrees x_i of a directed stochastic contact graph are independent and binomial (a_i, p) for $i \in \Omega$. If the sampling procedure is conditioned by fixed values $x_1, x_2, \ldots, x_N$ of the local out-degrees in the sample graph, the conditional sampling procedure will become identical with random sampling of x_i arcs without replacement from the set $\bigcup_j M_{ij}$ of out-arcs from node i in the population graph for $i \in \Omega$. Consequently, the various outcomes Z with $\sum_j Z_{ij} = x_i$ for $i \in \Omega$ will have the conditional probability

$$\prod_i \prod_j \binom{C_{ij}}{Z_{ij}} / \prod_i \binom{a_i}{x_i} . \tag{3}$$

1.2 Random choice of contacts

When arc sampling is considered as a procedure to obtain information about the population graph, it is assumed that the sampling procedure but not the population graph is completely specified. However, it is also possible to consider arc sampling from another point of view. The population graph may be completely specified, and the sampling procedure may be considered as a randomization device. If this device is known, the problem may be to derive probability distributions of some graph parameter. If the randomization device is incompletely known, the problem may be to use the stochastic contact graph to make inferences about the device or to estimate the probability distributions

of various graph parameters. These types of problem may conveniently be described in the following way.

The population graph is a known graph where the arcs represent possible contacts which may occur between the nodes. The occurrence of contacts is randomly determined and corresponds to a stochastic partial graph. If the contacts occur independently of each other and with a common occurrence probability p, the partial graph obtained is called a stochastic contact graph.

2. UNDIRECTED COMPLETE POPULATION GRAPHS

2.1 Introduction

Several problems which have been treated in sociometry and reliability theory can be expressed with the aid of stochastic contact graphs. In this section some problems will be described which are based on an undirected graph with possible contacts between all the node pairs and no multiple contacts, i.e.

$$A_{ij} = C_{ij} = \begin{cases} 1 & \text{if } i \neq j \ , \\ 0 & \text{if } i = j \ . \end{cases} \tag{4}$$

The stochastic contact graph can be represented by a symmetric arc indicator matrix X having $X_{ii} \equiv 0$ and X_{ij} independent Bernoulli (p) for $i < j$.

In Subsection 2.7 is discussed how the results of Subsections 2.2 - 2.6 are modified when circuit arcs are admitted, i.e. when $A_{ij} \equiv C_{ij} \equiv 1$.

2.2 Basic properties

The local degree x_i is binomial (N-1,p) for $i \in \Omega$. The local degrees x_i and x_j of two different nodes are dependent through X_{ij}. Obviously, the three variables X_{ij}, $x_i - X_{ij}$, and $x_j - X_{ij}$ are independent and Bernoulli (p), binomial (N-2,p), and binomial (N-2,p), respectively. Hence it follows that x_i and x_j for $i \neq j$ have the simultaneous probabilities

$$P(x_i=u,\ x_j=v) = q\ P(u,N-2)\ P(v,N-2) + p\ P(u-1,N-2)\ P(v-1,N-2) \tag{5}$$

for $u = 0,1,\ldots,N-1$ and $v = 0,1,\ldots,N-1$, with

$$P(u,U) = \binom{U}{u} p^u q^{U-u} \quad . \tag{6}$$

It also follows that

$$E\, x_i x_j = E\,[(x_i - X_{ij}) + X_{ij}][(x_j - X_{ij}) + X_{ij}] = N(N-2)p^2 + p \quad , \tag{7}$$

and consequently $\mathrm{Cov}\,(x_i, x_j) = pq$, and the coefficient of correlation between x_i and x_j is equal to $\rho(x_i, x_j) = 1/(N-1)$ for $i \neq j$.

The frequency $f_x(u)$ of nodes with the local degree equal to u satisfies the relations

$$E\, f_x(u) = \sum_{i=1}^{N} P(x_i = u) \quad , \tag{8}$$

$$\mathrm{Var}\, f_x(u) = \sum_{i \neq j} \sum P(x_i = u,\ x_j = u) + E\, f_x(u) - [E\, f_x(u)]^2 \quad , \tag{9}$$

$$\mathrm{Cov}\,[f_x(u), f_x(v)] = \sum_{i \neq j} \sum P(x_i = u,\ x_j = v) - E\, f_x(u)\, E\, f_x(v) \tag{10}$$

for $u \neq v$. Since x_i is binomial $(N-1, p)$ and the simultaneous probabilities are given by (5), the expected values, variances, and covariances of the distribution of the local degrees are obtained from (8), (9), and (10). Particularly, the number of isolates has the expected value

$$E\, f_x(0) = N\, q^{N-1} \quad , \tag{11}$$

and the variance

$$\mathrm{Var}\, f_x(0) = N(N-1)q^{2N-3} + N\, q^{N-1} - N^2 q^{2N-2} \quad . \tag{12}$$

2.3 Fixed number of contacts

Consider the stochastic contact graphs with a fixed number of contacts $\sum\sum_{i<j} X_{ij} = r$. Conditioned by r, there are $\binom{N(N-1)/2}{r}$ possible outcomes X which all have the same conditional probability. It follows that the conditional distribution of the local degree x_i is hypergeometric with

$$P(x_i=u \mid r) = \binom{N-1}{u}\binom{(N-1)(N-2)/2}{r-u}/\binom{N(N-1)/2}{r}, \tag{13}$$

and the conditional simultaneous distribution of x_i and x_j is given by

$$P(x_i=u,\ x_j=v \mid r) = \binom{1}{0}\binom{N-2}{u}\binom{N-2}{v}\binom{(N-2)(N-3)/2}{r-u-v}/\binom{N(N-1)/2}{r} +$$

$$+ \binom{1}{1}\binom{N-2}{u-1}\binom{N-2}{v-1}\binom{(N-2)(N-3)/2}{r-u-v+1}/\binom{N(N-1)/2}{r} \tag{14}$$

for $i \neq j$. If the formulas (8), (9), and (10) are applied to the conditional variables, the expressions (13) and (14) will provide the conditional expected values, variances, and covariances of the frequencies $f_x(u)$ for $u = 0,1,\ldots,N-1$.

2.4 The isolate problem of sociometry

The isolate problem of sociometry is to determine the probability distribution of the number of isolates, i.e. the probabilities

$$P(f_x(0) = n) \tag{15}$$

for $n = 0,1,\ldots,N$. The expected value and the variance of the number of isolates are given by (11) and (12). In order to determine the probabilities (15), the technique of compound probabilities can be used. This technique was founded by Fréchet (1940 and 1943) and is also described by Feller (1957). It is also referred to as the principle of inclusion and exclusion.

Consider the event that n fixed nodes $i_1, i_2, \ldots, i_n$ are isolated. There are $\binom{N}{n}$ such events, which all have the same probability $q^{\binom{N}{2}-\binom{N-n}{2}}$. The sum

of these probabilities is denoted

$$S_{Nn} = \binom{N}{n} q^{\binom{N}{2}-\binom{N-n}{2}} = \binom{N}{n} q^{N-1} q^{N-2} \dots q^{N-n} \tag{16}$$

for $n = 1,2,\dots,N$. Put $S_{N0} = 1$. According to Fréchet, it then holds that

$$P(f_x(0) = n) = \sum_{k=n}^{N} (-1)^{n+k} \binom{k}{n} S_{Nk} \tag{17}$$

for $n = 0,1,\dots,N$. Introduce the notation

$$P_N = P(f_x(0)=0) = \sum_{k=0}^{N} (-1)^k S_{Nk} \tag{18}$$

for the probability that there are no isolates in a stochastic contact graph with N nodes. The first instances of (18) are

$$\left.\begin{aligned} P_1 &= 0 \ , \\ P_2 &= 1-q \ , \\ P_3 &= 1-3q^2+2q^3 \ , \\ P_4 &= 1-4q^3+6q^5-3q^6 \ , \\ P_5 &= 1-5q^4+10q^7-10q^9+4q^{10} \ . \end{aligned}\right\} \tag{19}$$

The probability of exactly n isolates can be found from the probability that n specified nodes are isolates and no isolates exist among the other N-n nodes. By addition of such probabilities for all possible specifications of n nodes, it follows that

$$P(f_x(0)=n) = S_{Nn} P_{N-n} \tag{20}$$

for $n = 0,1,\dots,N$, if P_0 is defined as 1. The formula (20) can also be obtained by simple transformation of (17). According to (20), the desired probabilities

are obtained from the sums $S_{N0}, S_{N1}, \ldots, S_{NN}$ given by (16) and the probabilities $P_0, P_1, \ldots, P_N$ given by (18).

It may be convenient to have recourse to a recurrence formula to determine $P_0, P_1, \ldots, P_N$. Such a formula is obtained from (20), since the sum of the probabilities equals 1. It follows that

$$P_N = 1 - \sum_{n=1}^{N} S_{Nn} P_{N-n} = 1 - \sum_{k=0}^{N-1} S_{N,N-k} P_k \quad . \tag{21}$$

Simple lower and upper bounds of P_N can be obtained by application of general inequalities proved by Fréchet. For instance, it holds that

$$1 - N q^{N-1} \leq P_N \leq 1 - N q^{N-1} + \binom{N}{2} q^{2N-3} \quad . \tag{22}$$

2.5 The clique problem of sociometry

If the contact pattern between the individuals in a social group is represented by a complete graph, it is customary in sociometry to say that the individuals form a clique. This definition of a clique has been suggested by Festinger (1949). Luce and Perry (1949) and Luce (1950) have analyzed some clique variants algebraically without entering the probabilistic problems.

Consider all the $\binom{N}{n}$ subgraphs with n nodes which are present in a stochastic graph. The frequency of complete graphs among these subgraphs is denoted h_n. The clique problem of sociometry is to determine the probability distribution of the number of cliques, i.e. the distribution of h_n for $n \geq 2$. The difficulty of the problem suggests that one has to be satisfied with a determination of the expected value and the variance of the number of cliques.

Consider a fixed node subset with n nodes. The probability that the corresponding subgraph is complete equals $p^{\binom{n}{2}}$, and consequently

$$E\, h_n = \binom{N}{n} p^{\binom{n}{2}} \quad . \tag{23}$$

In order to calculate the variance of h_n, two fixed node subsets with n nodes each are considered. The probability that the corresponding two subgraphs are both complete equals $p^{2\binom{n}{2}-\binom{k}{2}}$, if the two node subsets have k nodes in common. Since the first node subset can be chosen in $\binom{N}{n}$ different ways, and since there are $\binom{n}{k}\binom{N-n}{n-k}$ possibilities of choosing the other node subset in such a way that k nodes belong to the first subset and n-k nodes do not, it follows that

$$E\, h_n^2 = \binom{N}{n} \sum_{k=0}^{n} \binom{n}{k}\binom{N-n}{n-k} p^{2\binom{n}{2}-\binom{k}{2}} . \tag{24}$$

Hence

$$\mathrm{Var}\, h_n = \binom{N}{n} \sum_{k=0}^{n} \binom{n}{k}\binom{N-n}{n-k} p^{2\binom{n}{2}-\binom{k}{2}} - \binom{N}{n}^2 p^{2\binom{n}{2}} . \tag{25}$$

Particularly, h_2 is equal to the arc frequency r in the stochastic contact graph. A simple check shows that for n=2, the expected value (23) and the variance (25) coincide with the corresponding values in a binomial $(\binom{N}{2}, p)$-distribution.

Moreover, h_3 is the number of three-node circuit paths in the stochastic contact graph. It follows from (23) and (25) that

$$E\, h_3 = \binom{N}{3} p^3 , \tag{26}$$

$$\mathrm{Var}\, h_3 = \binom{N}{3} p^3 + 3(N-3)\binom{N}{3} p^5 + \binom{N}{3}\left[3\binom{N-3}{2} + \binom{N-3}{3} - \binom{N}{3}\right] p^6 . \tag{27}$$

It may be mentioned that the results for cliques can be translated into results for node subsets consisting of n isolates. This is achieved by considering the complement of the occurred set of arcs, i.e. by interchanging p and q.

The conditional expected value and the conditional variance of h_n for a fixed arc frequency r can be obtained by an easy modification of the reasoning above. It holds that

$$E\,(h_n \mid r) = \binom{N}{n} \binom{\binom{N}{2}-\binom{n}{2}}{r-\binom{n}{2}} \Big/ \binom{\binom{N}{2}}{r}, \tag{28}$$

and

$$E\,(h_n^2 \mid r) = \binom{N}{n} \sum_{k=0}^{n} \binom{n}{k}\binom{N-n}{n-k} \binom{\binom{N}{2}-2\binom{n}{2}+\binom{k}{2}}{r-2\binom{n}{2}+\binom{k}{2}} \Big/ \binom{\binom{N}{2}}{r}. \tag{29}$$

2.6 Two problems of reliability

Certain reliability problems of interest in technical systems consisting of a large number of related components can be stated as probability problems concerning various types of connectedness in a graph. Gilbert (1959) has treated two such problems with applications to telephone traffic. The stochastic contact graph consists of N telephone central offices and the idle direct lines between them.

Using Gilbert's method, we will consider the probability P_N that the stochastic contact graph is connected and the probability Q_N that the stochastic graph has at least one path between two fixed nodes, say node 1 and node 2.

For a stochastic contact graph, one, and only one, of the N following events for $k=1,2,\ldots,N$ is true. Node 1 is connected to k-1 other nodes, and not one of these k connected nodes has any arcs to the other N-k nodes. Consequently it holds that

$$\sum_{k=1}^{N} \binom{N-1}{k-1} P_k\, q^{k(N-k)} = 1\ , \tag{30}$$

from which it follows that P_N can be reiteratively determined according to the formula

$$P_N = 1 - \sum_{k=1}^{N-1} \binom{N-1}{k-1} P_k\, q^{k(N-k)}\ , \tag{31}$$

with the initial value $P_1 = 1$. The first few values are

$$\left.\begin{aligned} P_2 &= 1-q \ , \\ P_3 &= 1-3q^2+2q^3 \ , \\ P_4 &= 1-4q^3-3q^4+12q^5-6q^6 \ , \\ P_5 &= 1-5q^4-10q^6+20q^7+30q^8-60q^9+24q^{10} \ . \end{aligned}\right\} \tag{32}$$

By use of generating functions, Gilbert obtains an explicit formula of P_N. This formula is too complicated to be applied in practice, and the reiterative formula is to be preferred.

The probability Q_N that there is a path between node 1 and node 2 can be determined in an analogous way. If node 1 and node 2 are not connected, one, and only one, of the N-1 following events for $k = 1, 2, \ldots, N-1$ is true. Node 1 is connected to k-1 nodes among $3, 4, \ldots, N$, and not one of these k connected nodes has any arcs to the other N-k nodes. Consequently it holds that

$$1 - Q_N = \sum_{k=1}^{N-1} \binom{N-2}{k-1} P_k \, q^{k(N-k)} \ . \tag{33}$$

Gilbert derives upper and lower bounds of P_N and Q_N, which are used to show that

$$P_N = 1 - N\,q^{N-1} + O(N^2 q^{3N/2}) \ , \tag{34}$$

and

$$Q_N = 1 - 2q^{N-1} + O(N\,q^{3N/2}) \tag{35}$$

for large values of N.

2.7 Modifications when circuit arcs are present

If the stochastic contact graph is represented by a symmetric arc indicator matrix X with X_{ij} independent Bernoulli (p) for $i \leq j$, i.e. if circuit arcs are also allowed, the results above are slightly modified.

The local degree x_i becomes binomial (N,p), and the total arc frequency r becomes binomial $(\binom{N+1}{2},p)$. The covariance between two different local degrees x_i and x_j is still equal to pq, and the coefficient of correlation $\rho(x_i,x_j)$ becomes equal to $1/N$ for $i \neq j$. The simultaneous probabilities are given by

$$P(x_i=u,\ x_j=v) = q\ P(u,N-1)\,P(v,N-1) + p\ P(u-1,N-1)\,P(v-1,N-1)\ , \qquad (36)$$

with $P(u,U)$ defined by (6). The treatment of the isolate problem in Subsection 2.4 is carried out in the same way, with (16) replaced by

$$S_{Nn} = \binom{N}{n} q^{\binom{N+1}{2}-\binom{N-n+1}{2}} . \qquad (37)$$

The number of cliques obtains the expected value

$$E\,h_n = \binom{N}{n} p^{\binom{n+1}{2}} , \qquad (38)$$

and the variance

$$\mathrm{Var}\,h_n = \binom{N}{n} \sum_{k=0}^{n} \binom{n}{k}\binom{N-n}{n-k} p^{2\binom{n+1}{2}-\binom{k+1}{2}} - \binom{N}{n}^2 p^{2\binom{n+1}{2}} . \qquad (39)$$

The reliability problem of Subsection 2.6 are obviously not influenced by circuit arcs.

3. DIRECTED COMPLETE POPULATION GRAPHS

3.1 Introduction

In sociometry it is usually the directed case that has been paid attention to. All the pairs of nodes are assumed to have possible contacts, i.e.

$$A_{ij} = C_{ij} = \begin{cases} 1 \text{ if } i \neq j \ , \\ 0 \text{ if } i = j \ . \end{cases} \tag{40}$$

The stochastic contact graph is represented by an arc indicator matrix X with $X_{ii} \equiv 0$ and X_{ij} independent Bernoulli (p) for $i \neq j$.

Usually it is assumed that the number of emitted contacts is fixed for each node, i.e. the local out-degrees $x_1, x_2, \ldots, x_N$ are fixed. The out-degrees are then also often assumed to be equal. The x_i contacts given by node i are assumed to be allocated at random among the N-1 other nodes. Different nodes are assumed to choose their emitted contacts independently of the choices of the other nodes. All these assumptions are realized if the stochastic contact graphs are conditioned by fixed values of the local out-degrees.

In this section some problems of the same kind as those of the previous section will be studied for directed graphs.

3.2 Basic properties

The local out-degrees x_i are independent binomial (N-1, p) for $i \in \Omega$. Also, the local in-degrees y_j are independent binomial (N-1, p) for $j \in \Omega$. At each node the local out-degree x_i and the local in-degree y_i are independent. For $i \neq j$, the local out-degree x_i and the local in-degree y_j are dependent through X_{ij}. The simultaneous distribution is given by

$$P(x_i=u, \ y_j=v) = q \ P(u, N-2) \ P(v, N-2) + p \ P(u-1, N-2) \ P(v-1, N-2) \ , \tag{41}$$

with P(u, U) defined by (6). In the same way as in Subsection 2.2 it follows that $\text{Cov}\ (x_i, y_j) = pq$ for $i \neq j$ and $\rho(x_i, y_j) = 1/(N-1)$ for $i \neq j$.

Since the local out-degrees are independent, it follows that the frequencies $[f_x(0), f_x(1), \ldots, f_x(N-1)]$ are multinomial $[N, P(0, N-1), P(1, N-1), \ldots, P(N-1, N-1)]$. Thus,

$$E \ f_x(u) = N \ P(u, N-1) \ , \tag{42}$$

$$\text{Var } f_x(u) = N\, P(u,N-1)\,[1-P(u,N-1)]\,, \tag{43}$$

$$\text{Cov}\,[f_x(u), f_x(v)] = -N\, P(u,N-1)\, P(v,N-1) \tag{44}$$

for $u \neq v$. Particularly, the number $f_x(0)$ of nodes without out-arcs is binomial (N, q^{N-1}).

Completely analogous results are true for the local in-degrees. When the local out-degrees and in-degrees are considered simultaneously, the distributional properties are more complex, due to the fact that the terms in the sum

$$f_{x,y}(u,v) = \sum_{i=1}^{N} I(x_i=u,\ y_i=v) \tag{45}$$

are not independent. However, it is straightforward to obtain the expected value

$$E\, f_{x,y}(u,v) = N\, P(u,N-1)\, P(v,N-1)\,, \tag{46}$$

and from the formula

$$\text{Var } f_{x,y}(u,v) = \sum_{i \neq j}\sum P(x_i=u,\ y_j=v)\, P(x_j=u,\ y_i=v) +$$

$$+ E\, f_{x,y}(u,v) - [E\, f_{x,y}(u,v)]^2 \tag{47}$$

the variance follows by substitution of (41). Particularly, the number of isolates has the expected value

$$E\, f_{x,y}(0,0) = N\, q^{2N-2}\,, \tag{48}$$

and the variance

$$\text{Var } f_{x,y}(0,0) = N(N-1)\, q^{4N-6} + N\, q^{2N-2} - N^2\, q^{4N-4}\,. \tag{49}$$

3.3 Fixed number of contacts

When the total number r of contacts is fixed, all the $\binom{N(N-1)}{r}$ possible outcomes X have the same probability. The local out-degrees $x_1, x_2, \ldots, x_N$ have a conditional distribution given by

$$P(x_1=u_1, \ldots, x_N=u_N \mid r) = \prod_{i=1}^{N} \binom{N-1}{u_i} \Big/ \binom{N(N-1)}{r} . \tag{50}$$

The unidimensional and two-dimensional marginal distributions are given by

$$P(x_i=u \mid r) = \binom{N-1}{u}\binom{(N-1)^2}{r-u} \Big/ \binom{N(N-1)}{r} , \tag{51}$$

and

$$P(x_i=u,\ x_j=v \mid r) = \binom{N-1}{u}\binom{N-1}{v}\binom{(N-1)(N-2)}{r-u-v} \Big/ \binom{N(N-1)}{r} \tag{52}$$

for $i \neq j$. The local in-degrees $y_1, y_2, \ldots, y_N$ have the same conditional distributions. The simultaneous conditional distributions of the local out-degrees and in-degrees are more involved. The simplest simultaneous distributions are given by

$$P(x_i=u,\ y_i=v \mid r) = \binom{N-1}{u}\binom{N-1}{v}\binom{(N-1)(N-2)}{r-u-v} \Big/ \binom{N(N-1)}{r} , \tag{53}$$

and

$$P(x_i=u,\ y_j=v \mid r) = \binom{1}{0}\binom{N-2}{u}\binom{N-2}{v}\binom{N^2-3N+3}{r-u-v} \Big/ \binom{N(N-1)}{r} +$$

$$+ \binom{1}{1}\binom{N-2}{u-1}\binom{N-2}{v-1}\binom{N^2-3N+3}{r-u-v+1} \Big/ \binom{N(N-1)}{r} \tag{54}$$

for $i \neq j$.

3.4 Fixed local out-degrees

When the local out-degrees $x_1, x_2, \ldots, x_N$ are fixed, all the $\binom{N-1}{x_1}\binom{N-1}{x_2}\ldots\binom{N-1}{x_N}$ possible outcomes X have the same probability. Conditioned by $x_1, x_2, \ldots, x_N$, the rows of X are stochastically independent. Hence the local in-degree $y_j = \sum_{i=1}^{N} X_{ij}$ is a sum of N-1 independent Bernoulli $(x_i/(N-1))$ variables with $i \neq j$. Particularly, when all the out-degrees are equal, say $x_i \equiv k$, it follows that the conditional distribution of y_j is binomial $(N-1, k/(N-1))$, which for large values of N may be approximated by a Poisson (k) distribution.

In order to determine the simultaneous probabilities

$$P(y_i=u,\ y_j=v \mid x_1=\ldots=x_N=k) \tag{55}$$

for $i \neq j$, four different cases have to be distinguished, corresponding to the various combinations of values of X_{ij} and X_{ji}. Put for the sake of brevity

$$\left.\begin{aligned} p_{00} &= (N-k-1)(N-k-2)/(N-1)(N-2)\ , \\ p_{01} &= p_{10} = k(N-k-1)/(N-1)(N-2)\ , \\ p_{11} &= k(k-1)/(N-1)(N-2)\ , \end{aligned}\right\} \tag{56}$$

and

$$\binom{N}{u,v,w} = N!\, p_{11}^{w} p_{10}^{u-w} p_{01}^{v-w} p_{00}^{N-u-v+w}/w!(u-w)!(v-w)!(N-u-v+w)!. \tag{57}$$

It is possible to interpret (57) as the probability that among N stochastically independent pairs of binary digits there are u with the digit 1 in the first position, v with the digit 1 in the second position, and w with the digit 1 in both positions, i.e. the outcomes 11, 10, 01, and 00 have the frequencies w, u-w, v-w, and N-u-v+w, respectively. The outcomes 11, 10, 01, and 00 are supposed to have the probabilities p_{11}, p_{10}, p_{01}, and p_{00}, respectively.

Consider the columns i and j of X. If w denotes the number of rows of X which have the digit 1 in both the columns i and j, the above interpretation of (57) may be applied to obtain

$$P(y_i=u,\ y_j=v \mid x_1=\ldots=x_N=k) =$$

$$= \sum_{w=0}^{N-2} \left[p_0^2 \binom{N-2}{u,v,w} + p_0 p_1 \binom{N-2}{u,v-1,w} + p_1 p_0 \binom{N-2}{u-1,v,w} + \right.$$

$$\left. + p_1^2 \binom{N-2}{u-1,v-1,w} \right] , \tag{58}$$

where $p_0=p_{00}+p_{01}$ and $p_1=p_{10}+p_{11}$ are the probabilities that X_{ij} (or X_{ji}) equals 0 and 1, respectively.

3.5 The reciprocal choice problem of sociometry

The reciprocal choice problem of sociometry is to determine the probability distribution of the number of pairs of nodes which mutually contact each other. The number of reciprocal choices can be written

$$t = \sum_{i<j} \sum X_{ij} X_{ji} , \tag{59}$$

and it is binomial $\left(\binom{N}{2}, p^2\right)$ in the unconditional case. If there is a fixed number of contacts, the conditional probabilities $P(t=k \mid r)$ can be derived from a consideration of the $\binom{N}{2}$ pairs (X_{ij}, X_{ji}) for $i<j$. There will be k pairs 11, r-2k pairs 10 or 01, and $\binom{N}{2}$-r+k pairs 00. The r-2k pairs 10 or 01 can be selected in 2^{r-2k} different ways. The order between the pairs can be varied in

$$\binom{N}{2}!/k!(r-2k)!\left[\binom{N}{2}-r+k\right]! \tag{60}$$

different ways. It follows that

$$P(t=k \mid r) = \binom{N}{2}!\ 2^{r-2k}/k!(r-2k)!\left[\binom{N}{2}-r+k\right]!\binom{N(N-1)}{r} \tag{61}$$

for all non-negative integers k satisfying

$$\left.\begin{aligned} &0 \le k \le \binom{N}{2} , \\ &0 \le r-2k \le \binom{N}{2} , \\ &0 \le \binom{N}{2}-r+k \le \binom{N}{2} , \end{aligned}\right\} \tag{62}$$

i.e. non-negative integers k satisfying

$$[r-\binom{N}{2}]/2 \le k \le r/2 . \tag{63}$$

The conditional expected value of t is most easily obtained from (59). Thus,

$$E(t \mid r) = \binom{N}{2}\binom{r}{2}/\binom{N(N-1)}{2} = \binom{r}{2}/(N^2-N-1) . \tag{64}$$

Moreover,

$$\mathrm{Var}(t \mid r) = \binom{N}{2}\binom{r}{2}/\binom{N(N-1)}{2} + \binom{N}{2}[\binom{N}{2}-1]\binom{r}{4}/\binom{N(N-1)}{4} -$$

$$- [E(t \mid r)]^2 . \tag{65}$$

If r and N tend to infinity in such a way that $r/N(N-1)$ approaches a fixed limit p, it follows from (64) and (65) that $E(t \mid r)/\binom{N}{2}$ approaches p^2 and $[\mathrm{Var}(t \mid r)]/\binom{N}{2}$ approaches $p^2(1-p^2)$, which is in accordance with the binomial distribution obtained above in the unconditional case.

Let us now turn to the case of fixed local out-degrees $x_1, x_2, \ldots, x_N$. The conditional probability distribution of the reciprocal choice frequency t for fixed local out-degrees is the case most commonly dealt with in sociometry. According to (59) and the fact that the rows of X are stochastically independent, it follows that

$$E(t \mid x_1, \ldots, x_N) = \sum_{i<j} \sum x_i x_j/(N-1)^2 . \tag{66}$$

If the mean value and the standard deviation of the local out-degrees are intro-

duced, the result can also be written

$$E\,(t \mid x_1, \ldots, x_N) = N\, m^2(x)/2(N-1) - N\, s^2(x)/2(N-1)^2 \quad . \tag{67}$$

The formula

$$\text{Var } t = \sum_{i \neq j} \sum \sum_{k \neq l} \text{Cov}\,(X_{ij}X_{ji}, X_{kl}X_{lk})/4 =$$

$$= \sum_{i \neq j} \sum [\text{Var}\,(X_{ij}X_{ji})]/2 + \sum_{i \neq j \neq k \neq i} \sum \sum \text{Cov}\,(X_{ij}X_{ji}, X_{ik}X_{ki}) \quad , \tag{68}$$

where all the variables are understood to be conditioned by $x_1, x_2, \ldots, x_N$, gives the following conditional variance:

$$\text{Var}\,(t \mid x_1, \ldots, x_N) = \frac{1}{2} \sum_{i \neq j} \sum \left[1 - \frac{x_i x_j}{(N-1)^2}\right] +$$

$$+ \sum_{i \neq j \neq k \neq i} \sum \sum \left[\frac{x_i(x_i-1)x_j x_k}{(N-1)^3(N-2)} - \frac{x_i^2 x_j x_k}{(N-1)^4}\right] \quad . \tag{69}$$

Particularly, it follows from (67) and (69) that in the case of equal local out-degrees the expected value becomes

$$E\,(t \mid x_1 = \ldots = x_N = k) = N\, k^2/2(N-1) \quad , \tag{70}$$

and the variance

$$\text{Var}\,(t \mid x_1 = \ldots = x_N = k) = N\, k^2 (N-k-1)^2/2(N-1)^3 \quad . \tag{71}$$

The formula (70) for the expected number of reciprocal choices when each node emits k contacts has been given by several authors, for instance Moreno and Jennings (1938), Bronfenbrenner (1943), Criswell (1947), and Edwards (1948). The formula (71) for the variance has been derived by Weibull (1958) by

use of the Fréchet technique of compound probabilities. In principle Weibull has also given a variance formula for the case of unequal local out-degrees. However, his formula is indeed very complicated, since it is derived and expressed in terms of the frequencies $f_x(u)$ for $u = 0, 1, \ldots, N-1$.

If we put $k/(N-1) = p$, the expected value in (70) becomes $\binom{N}{2}p^2$, and the variance in (71) becomes $\binom{N}{2}p^2(1-p)^2$. If these expressions are compared with the unconditional formulas for the binomial $(\binom{N}{2}, p^2)$ distribution, it will be found that the expected values are equal, but the variance is smaller in the conditional than in the unconditional case.

It is not likely that it is possible to provide a useful general formula for the conditional probability distribution of the number of reciprocal choices t for fixed values of the local out-degrees $x_1, x_2, \ldots, x_N$. Such a formula does not seem accessible in the case of equal local out-degrees, either. However, the probabilities may in principle be obtained by use of the Fréchet technique of compound probabilities. This is demonstrated by Weibull (1958), who gives formulas which make it possible to compute the conditional probabilities in the case of equal local out-degrees and at most five nodes in the graph. When the number of nodes in the graph approaches infinity, Weibull shows that the conditional probability distribution of $(t \mid x_1 = \ldots = x_N = k)$ converges towards a Poisson $(k^2/2)$ distribution.

3.6 The isolate problem of sociometry

In Subsection 3.2 the frequency $f_x(0)$ of nodes without out-arcs was shown to be binomial (N, q^{N-1}). The same distribution is true for the frequency $f_y(0)$ of nodes without in-arcs. For the frequency $f_{x,y}(0,0)$ of nodes without arcs, the expected value and the variance have been given in (48) and (49).

The conditional expected value

$$E[f_x(0) \mid r] = \sum_{i=1}^{N} P(x_i = 0 \mid r) \tag{72}$$

and the conditional variance

$$\mathrm{Var}\,[f_x(0) \mid r] = \sum_{i \neq j} \sum P(x_i=0,\ x_j=0 \mid r) + E\,[f_x(0) \mid r] - \left[E\,[f_x(0) \mid r]\right]^2 \quad (73)$$

are obtained by use of the formulas (51) and (52). The conditional expected value and variance of $f_{x,y}(0,0)$ can be obtained in an analogous way.

The most common conditioning used in sociometry is fixed local out-degrees. Then the isolate problem is to determine the conditional probabilities

$$P(f_y(0) = n \mid x_1, \ldots, x_N) \quad (74)$$

for $n = 0, 1, \ldots, N$. This problem has been solved by Katz (1951) in the particular case of equal local out-degrees $x_1 = \ldots = x_N = k$. The general case has been treated by Katz and Powell (1957), who provide a solution based upon the determination of the number of matrices X which have fixed row and column sums. This combinatorial problem has been studied by Katz and Powell (1954).

The particular case $x_1 = \ldots = x_N = k$ was investigated by Katz (1951) by use of the Fréchet technique of compound probabilities. This technique is also applicable to the general case, which will now be demonstrated.

Consider an arbitrary node subset $S \subset \Omega$ with n nodes. The probability that the n nodes of S are isolated, i.e. have no in-arcs, is equal to

$$\prod_{i \in S} \binom{N-n}{x_i} \prod_{j \in \bar{S}} \binom{N-n-1}{x_j} \Big/ \prod_{i \in \Omega} \binom{N-1}{x_i} . \quad (75)$$

By use of the notations

$$c_n = \prod_{i \in \Omega} \left[\binom{N-n}{x_i} \Big/ \binom{N-1}{x_i} \right] \quad (76)$$

and

$$q_n(j) = (N-n-x_j)/(N-n) , \quad (77)$$

the probability (75) can be written

$$c_n \prod_{j \in \bar{S}} q_n(j) \ . \tag{78}$$

It should be noted that $c_n = 0$ if any of the local out-degrees is larger than N-n. If $x_j = N-n$, then $q_n(j) = 0$. As a consequence of this it follows that (78) attains the value zero correctly according to (75). Let S_n denote the sum of the $\binom{N}{n}$ probabilities (78) which belong to the different node subsets with n nodes. Then

$$S_n = c_n \sum_{j_1 < \ldots < j_{N-n}} \ldots \sum q_n(j_1) \ldots q_n(j_{N-n}) \ , \tag{79}$$

and according to Fréchet,

$$P(f_y(0)=n \mid x_1, \ldots, x_N) = \sum_{m=n}^{N} (-1)^{n+m} \binom{m}{n} S_m \tag{80}$$

for $n = 1, 2, \ldots, N$. If S_0 is defined as 1, the formula (80) is true also for $n = 0$.

In the particular case $n = 1$, the formula (79) can be written

$$S_1 = \sum_{j_1 < \ldots < j_{N-1}} \ldots \sum q_1(j_1) \ldots q_1(j_{N-1}) = q_1(1) \ldots q_1(N) \sum_{i=1}^{N} 1/q_1(i) \ , \tag{81}$$

where the last expression is well-defined if $q_1(i) > 0$ for $i = 1, 2, \ldots, N$. In the particular case $n = 2$, it holds that

$$S_2 = q_1(1) \ldots q_1(N) \sum_{j_1 < \ldots < j_{N-2}} \ldots \sum q_2(j_1) \ldots q_2(j_{N-2}) =$$

$$= q_1(1) \ldots q_1(N) q_2(1) \ldots q_2(N) \sum_{i<j} \sum 1/q_2(i) q_2(j) \ , \tag{82}$$

where the last expression is well-defined if $q_1(i) q_2(i) > 0$ for $i = 1, 2, \ldots, N$.

Fréchet (1940 and 1943) has shown that $n!\, S_n$ equals the n-th factorial mo-

ment of the distribution of $f_y(0)$. Hence

$$E\left[f_y(0) \mid x_1, \ldots, x_N\right] = S_1 \ , \tag{83}$$

and

$$\mathrm{Var}\left[f_y(0) \mid x_1, \ldots, x_N\right] = 2\,S_2 + S_1 - S_1^2 \ , \tag{84}$$

where S_1 and S_2 are given by (81) and (82).

In the particular case of equal local out-degrees $x_1 = \ldots = x_N = k$, it follows from (79) that

$$S_n = \binom{N}{n}\binom{N-n}{k}^n \binom{N-n-1}{k}^{N-n} \Big/ \binom{N-1}{k}^N \tag{85}$$

for $n = 0, 1, \ldots, N-k-1$, and $S_n = 0$ otherwise. Put $q_1 = 1-k/(N-1)$ and $q_2 = 1-k/(N-2)$. Then,

$$E\left[f_y(0) \mid x_1 = \ldots = x_N = k\right] = N\,q_1^{N-1} \ , \tag{86}$$

and

$$\mathrm{Var}\left[f_y(0) \mid x_1 = \ldots = x_N = k\right] = N(N-1)q_1^N q_2^{N-2} + N\,q_1^{N-1} - N^2 q_1^{2N-2} . \tag{87}$$

For large values of N these results are in close agreement with the results obtained from the binomial (N, q^{N-1}) distribution which occurs in the unconditional case where q is approximately equal to q_1 and q_2. These approximate properties have been examined in more detail by Katz (1951).

3.7 The clique problem of sociometry

The results of Subsection 2.5 concerning the clique problem in the undirected case are easily modified to fit the directed case.

If h_n is the number of node subsets with n nodes which correspond to complete subgraphs of the stochastic contact graph, it follows exactly as in the undirected case that

$$E\, h_n = \binom{N}{n} p^{n(n-1)} , \tag{88}$$

and

$$\mathrm{Var}\, h_n = \binom{N}{n} \sum_{k=0}^{n} \binom{n}{k}\binom{N-n}{n-k} p^{2n(n-1)-k(k-1)} - (E\, h_n)^2 . \tag{89}$$

If the total arc frequency r is fixed, the corresponding conditional values are

$$E\,(h_n \mid r) = \binom{N}{n}\binom{N(N-1)-n(n-1)}{r-n(n-1)} \Big/ \binom{N(N-1)}{r} , \tag{90}$$

and

$$\mathrm{Var}\,(h_n \mid r) = \binom{N}{n} \sum_{k=0}^{n} \binom{n}{k}\binom{N-n}{n-k}\binom{N(N-1)-2n(n-1)+k(k-1)}{r-2n(n-1)+k(k-1)} \Big/ \binom{N(N-1)}{r} -$$

$$- [E\,(h_n \mid r)]^2 . \tag{91}$$

Finally consider the case of fixed local out-degrees. Let S be an arbitrary node subset with n nodes. Then

$$E\,(h_n \mid x_1, \ldots, x_N) = \sum_S \prod_{i \in S} \left[\binom{N-n}{x_i - n+1} \Big/ \binom{N-1}{x_i} \right] =$$

$$= \sum_S \prod_{i \in S} \binom{x_i}{n-1} \Big/ \binom{N-1}{n-1}^n = \sum_{i_1 < \ldots < i_n} \cdots \sum \binom{x_{i_1}}{n-1} \cdots \binom{x_{i_n}}{n-1} \Big/ \binom{N-1}{n-1}^n . \tag{92}$$

Particularly, h_2 is equal to the number of reciprocal choices t dealt with in Subsection 3.5. A simple check shows that for n = 2, (92) agrees with (66). If

S_1 and S_2 are two node subsets with n nodes each, it follows that

$$E\,(h_n^2 \mid x_1,\ldots,x_N) = \sum_{S_1}\sum_{S_2}\prod_i\left[\binom{x_i}{n-1}/\binom{N-1}{n-1}\right]\prod_j\left[\binom{x_j}{2n-m-1}/\binom{N-1}{2n-m-1}\right], \quad (93)$$

where $i \in (S_1 \cap \bar{S}_2) \cup (\bar{S}_1 \cap S_2)$, $j \in S_1 \cap S_2$, and m is the number of nodes in $S_1 \cap S_2$. Particularly, it follows from (92) and (93) that

$$E\,(h_n \mid x_1=\ldots=x_N=k) = \binom{N}{n}\binom{k}{n-1}^n/\binom{N-1}{n-1}^n , \quad (94)$$

and

$$E\,(h_n^2 \mid x_1=\ldots=x_N=k) =$$

$$= \binom{N}{n}\sum_{m=0}^{n}\binom{n}{m}\binom{N-n}{n-m}\left[\binom{k}{n-1}/\binom{N-1}{n-1}\right]^{2(n-m)}\left[\binom{k}{2n-m-1}/\binom{N-1}{2n-m-1}\right]^m . \quad (95)$$

3.8 Modifications when circuit arcs are present

If the stochastic contact graph is represented by arc indicator variables X_{ij} which are independent and Bernoulli (p) for all i and j, i.e. if circuit arcs may also occur, the results given above in this section are slightly modified. The modifications are of the same kind as in Subsection 2.7, and no essentially different problems are involved. The details are therefore omitted.

4. GENERAL POPULATION GRAPHS

4.1 The undirected case

The undirected stochastic contact graphs which were treated in Section 2 had a stochastic arc between every pair of nodes. This case may be characterized as arc sampling from a complete undirected population graph. We will now turn to general undirected population graphs which are represented by a symmetric

arc indicator matrix A.

The stochastic contact graph will be represented by a symmetric arc indicator matrix X with $X_{ij} = 0$ if $A_{ij} = 0$, and X_{ij} independent Bernoulli (p) if $A_{ij} = 1$ for $i \leq j$.

If A is a known matrix, this is a direct generalization of the case treated in Section 2. If A is unknown, the stochastic contact graph may be considered as a sample graph, which will be used to make inferences on the population graph.

4.2 Estimation of the distribution of the local degrees

The local degree x_i of node i in the sample graph is binomial (a_i, p). The local degrees x_i and x_j of two different nodes i and j are independent binomial (a_i, p) and binomial (a_j, p) if $A_{ij} = 0$. When $A_{ij} = 1$, the local degrees x_i and x_j become dependent through X_{ij}. Then X_{ij}, $x_i - X_{ij}$, and $x_j - X_{ij}$ are independent Bernoulli (p), binomial $(a_i - 1, p)$, and binomial $(a_j - 1, p)$, respectively. Consequently,

$$P(x_i = u,\ x_j = v) = (1 - A_{ij})\, P(u, a_i)\, P(v, a_j) +$$

$$+ A_{ij} \left[q\, P(u, a_i - 1)\, P(v, a_j - 1) + p\, P(u-1,\ a_i - 1)\, P(v-1,\ a_j - 1) \right] \qquad (96)$$

for $i \neq j$, where $P(u, U)$ is defined by (6).

These results can be used to determine the expected values, the variances, and the covariances of the frequencies $f_x(u)$ for $u = 0, 1, \ldots, N$. According to the formula (8), it follows that

$$E\, f_x(u) = \sum_{i=1}^{N} P(u, a_i) = \sum_{U=0}^{N} P(u, U) f_a(U) \ , \qquad (97)$$

and the variances and covariances are obtained according to the formulas (9), (10), (96), and (97). In the same way as in Chapters 4 and 5, the relation (97) for $u = 0, 1, \ldots, N$ implies that it is possible to obtain unbiased estimators $\hat{f}_a(U)$ of $f_a(U)$ as the solutions of the triangular equation system

$$f_x(u) = \sum_{U=u}^{N} P(u,U)\, \hat{f}_a(U) \tag{98}$$

for $u = 0,1,\ldots,N$. It follows that

$$\hat{f}_a(U) = \sum_{u=U}^{N} Q(U,u)\, f_x(u) \tag{99}$$

for $U = 0,1,\ldots,N$, with

$$Q(U,u) = \binom{u}{U} (1/p)^U (1-1/p)^{u-U} \ . \tag{100}$$

The variance of the estimator $\hat{f}_a(U)$ can be expressed in terms of the covariances Cov $[f_x(u), f_x(v)]$ obtainable through the formulas (10) and (96). In order to obtain estimators of Var $\hat{f}_a(U)$, it seems likely that the simultaneous distribution of (a_i, a_j, A_{ij}) has to be introduced and estimated.

4.3 The isolate problem

Consider the problem of determining the probabilities $P(f_x(0) = n)$ for $n = 0,1,\ldots,N$. The technique of Subsection 2.4 can be applied if (16) is replaced by

$$S_{Nn} = \sum_{S} q^{R(S)} \ , \tag{101}$$

where the sum is carried out for all the $\binom{N}{n}$ node subsets $S \subset \Omega$ which consist of n nodes, and where R(S) denotes the arc frequency in the corresponding subgraph of the population graph, i.e.

$$R(S) = \sum_{i \le j} \sum A_{ij}\, I(i \in S,\ j \in S) \ . \tag{102}$$

If the number of subgraphs of the population graph which have n nodes and m

arcs is denoted by F_{nm}, it is possible to write (101) as

$$S_{Nn} = \sum_m F_{nm} q^m . \tag{103}$$

Then in the ordinary way,

$$P(f_x(0)=n) = \sum_{k=n}^{N} (-1)^{n+k} \binom{k}{n} S_{Nk} = \sum_m q^m \sum_{k=n}^{N} (-1)^{n+k} \binom{k}{n} F_{km} \tag{104}$$

for $n = 1, 2, \ldots, N$, and

$$P(f_x(0)=0) = 1 + \sum_m q^m \sum_{k=1}^{N} (-1)^k F_{km} . \tag{105}$$

4.4 An application to quality control

Imagine that quality control in the production of grammophone records will be performed in the following way. From a population set of grammophone records a certain number of records are randomly selected, and some sort of test procedure applied to both sides of the sampled records.

Each side of a grammophone record is said to contain one musical item. Each record has two different musical items, and the same musical item may occur on several records. All the records of the population are different, i.e. any two records may have at most one musical item in common.

How many times is it expected that the different musical items will be tested? What is the probability that each musical item is tested at least once? How many musical items are expected to be tested more than once? These and similar questions can be stated by use of graph terminology in the following way.

Introduce an undirected graph with N nodes corresponding to the different musical items, and R arcs corresponding to the grammophone records in the population. Each arc connects the two nodes which correspond to the two musical items of the record. The local degree a_i is equal to the number of records with the i-th musical item. The number of musical items which occur on exactly U records

equals $f_a(U)$.

The sampled grammophone records are represented by a partial graph which is given by a matrix X. The local degree x_i is equal to the number of tests performed on the i-th musical item. The number of musical items that are tested exactly u times is equal to $f_x(u)$.

The questions stated above may be answered in the following way. It is expected that the i-th musical item will be tested E x_i times. The probability that each musical item is tested at least once equals

$$P(x_i > 0 \text{ for } i = 1, 2, \ldots, N) = P(f_x(0) = 0) \ . \tag{106}$$

The number of musical items which may be expected to be tested more than once equals

$$E \sum_{u=2}^{N} f_x(u) = N - E\, f_x(0) - E\, f_x(1) \ . \tag{107}$$

In connection with this application it is easy to imagine other kinds of problem which may be treated by graph methods. For instance, problems of production planning may be considered where the purpose is to allocate musical items on the records in a way that is optimal according to some criterion. Also, inference problems may be considered where the purpose is to use the information provided by tests on some sampled grammophone records to estimate various properties of the musical item records in a population.

4.5 The directed case

Consider a directed population graph represented by an arc indicator matrix A. The stochastic contact graph is represented by an arc indicator matrix X, with $X_{ij} = 0$ if $A_{ij} = 0$, and X_{ij} independent Bernoulli (p) if $A_{ij} = 1$. If A is a known matrix, the stochastic contact graph is a direct generalization of the case treated in Section 3. If A is unknown, the stochastic contact graph will be considered as a sample graph which is to be used to make inferences on the popula-

tion graph.

The local out-degrees x_i are independent binomial (a_i, p) for $i = 1, 2, \ldots, N$. In the same way, the local in-degrees y_j are independent binomial (b_j, p) for $j = 1, 2, \ldots, N$. The total arc frequency r of the stochastic contact graph is binomial (R, p). If $A_{ij} = 0$, it follows that x_i and y_j are independent. If $A_{ij} = 1$, x_i and y_j are dependent through X_{ij}, and in fact X_{ij}, $x_i - X_{ij}$, and $y_j - X_{ij}$ are independent Bernoulli (p), binomial $(a_i - 1, p)$ and binomial $(b_j - 1, p)$, respectively.

The frequency $f_x(u)$ of nodes with local out-degree equal to u will satisfy

$$E\, f_x(u) = \sum_{U=0}^{N} P(u, U) f_a(U) \quad , \tag{108}$$

$$\mathrm{Var}\, f_x(u) = \sum_{U=0}^{N} P(u, U)\,[1 - P(u, U)]\, f_a(U) \quad , \tag{109}$$

$$\mathrm{Cov}\,[f_x(u), f_x(v)] = -\sum_{U=0}^{N} P(u, U)\, P(v, U)\, f_a(U) \tag{110}$$

for $u \neq v$. In the simple case of equal local out-degrees $a_1 = a_2 = \ldots = a_N = U$, the frequencies $[f_x(0), f_x(1), \ldots, f_x(U)]$ are multinomial $[N, P(0, U), P(1, U), \ldots, P(U, U)]$. The local in-degrees have analogous properties. The simultaneous frequencies $f_{x,y}(u, v)$ are easily handled, if there are no circuit arcs in the population graph. In fact, if $A_{ii} \equiv 0$, the expected values, variances, and covariances are obtainable in the ordinary way by use of the probabilities

$$P(x_i = u,\ y_i = v,\ x_j = u',\ y_j = v') = P(x_i = u,\ y_j = v')\, P(x_j = u',\ y_j = v) \quad , \tag{111}$$

and

$$P(x_i = u,\ y_j = v) = (1 - A_{ij})\, P(u, a_i)\, P(v, b_j) +$$

$$+ A_{ij}\,[q\, P(u, a_i - 1)\, P(v, b_j - 1) + p\, P(u - 1,\ a_i - 1)\, P(v - 1,\ b_j - 1)] \tag{112}$$

for $i \neq j$. If circuit arcs are present in the population graph, the variance and

covariance formulas become considerably more complicated.

Unbiased estimators $\hat{f}_a(U)$ of $f_a(U)$ for $U = 0, 1, \ldots, N$ can be obtained from the equation system (98), with the solutions given by (99) and (100). In the directed case it is easy to determine the variances of the estimators and to estimate these variances unbiasedly by use of the frequencies $f_x(u)$ for $u = 0, 1, \ldots, N$. The procedure is completely analogous to the one demonstrated in Subsection 6.4 of Chapter 4.

Chapter 8

INFERENCE FROM STOCHASTIC PREFERENCE GRAPHS

SYNOPSIS

1. Ranking based on paired comparisons. Various occasions when ranking of objects is performed are exemplified, and the experimental method of paired comparisons is discussed. Comparison graphs and preference graphs are introduced. Methods for testing the existence of an order and for ranking the objects are considered.

2. Stochastic preference graphs. Ranking by use of preference probabilities is discussed. A special stochastic preference model is defined and illustrated by some different experimental situations where the model is applicable.

3. Complete preference graphs. Each pair of objects is assumed to be judged and preference ordered a constant number of times in accordance with the given model. Stochastic properties are studied, and generalizations are provided of results by Kendall and Smith concerning the agreement between the preferences of different judges and the occurrence of inconsistencies among the

preferences. The probability of obtaining consistent preferences is given. Maximum likelihood estimation of the preference probability is discussed.

4. Complete bipartite preference graphs. There are two sets of objects; between the sets all possible comparisons are performed, whereas no comparison is made within the sets. Agreement is studied in a way analogous to that in Section 3. The number of circuits with four nodes is used as a measure of inconsistency, and the mean and variance properties are studied.

1. RANKING BASED ON PAIRED COMPARISONS

1.1 Paired comparisons and preferences

In several connections there is a need to apply what might be called experimental methods of judging and ranking a lot of objects. The objects to be compared may be the players in a tournament, different drugs in a test, various brands in a study of consumers' behavior, etc. Quite generally, the methods of ranking occur in connection with various kinds of priority scheduling, preference order, and choice behavior. Methods of ranking have deserved special attention in psychometric applications; see for instance Guilford (1954). Torgerson (1958), and Restle (1961). The economic theory of utility and the theory of subjective probability contain ranking as a basic element (Kyburg and Smokler, 1964).

The most common method of ranking is based on paired comparisons between the objects which are to be judged. A monograph by David (1963) describes the various models which have been suggested in the literature for studying paired comparisons. The method of paired comparisons usually presupposes that all the paired comparisons are performed, – this method is experimentally rather comprehensive and time consuming. If each one of n judges is to compare every pair among N objects, a total of $\binom{N}{2}n$ comparisons have to be carried out.

Paired comparison experiments are mainly performed in such a way that the objects are subjectively judged, i.e. it is impossible or impracticable to base the comparisons upon physical measurements. It is possible to distinguish cases where a prior order is known to exist between the objects, and cases where such an order need not exist.

The get some objects ordered according to decreasing weights by lifting them two and two in the hands is an example of a case where a prior order exists but may be hart to find, due to the imperfectness of the paired comparison procedure, i.e. the lifting. Experiments of this kind may be used to study consistency and other properties of the choice behavior. It is then principally the ranking ability of the judge that is investigated.

If some objects are to be graded according to decreasing beauty, or if some persons are to be ranked according to decreasing skill, it is obviously not certain that such a ranking can be performed, since the concepts of beauty and skill may be defined in such a way that ranking becomes impossible. In such a case, the inconsistencies that may occur among the preferences can be the result of either the insufficient ability of the judge or the lack of an order between the objects.

If the comparisons between the objects can be based upon several characteristics, the existence of an order cannot be taken for granted. As an example assume that a person tries to state his preferences for a number of employments by considering their respective wage and work satisfaction to him. If the employments with the higher wages also provide the higher satisfaction, they can be preference ordered. If, on the contrary, the employments with the higher wages offer a lower work satisfaction, it is possible to get opposite preference orders depending on whether it is the wages or the work satisfaction that is most highly valued. Generally, the existence of a preference order in this case implies that the wage and the work satisfaction can be simultaneously valued.

Another case where no order may exist is when the outcomes of the paired comparisons are settled by the majority in a collective. Suppose that the objects 1, 2, and 3 are ranked 1, 2, 3 by one person, 2, 3, 1 by another and 3, 1, 2 by a third person. If these three people vote for all paired comparisons and agree to the majority preference as the collective preference, no collective ranking will exist. The object 1 will be collectively preferred to object 2, which will be collectively preferred to object 3, which will be collectively preferred to object 1.

1.2 Comparison graphs and preference graphs

The paired comparisons can be represented by an undirected comparison graph. The objects to be compared are represented by the nodes and the comparisons by the arcs of the graph. The outcomes of the comparisons can be indicated by orientation of the arcs. A directed arc from node i to node j is taken to mean that object i is preferred to object j. In this way, a directed preference

graph is obtained from the undirected comparison graph.

If the obtained preference graph has no circuit path, it represents a partial order between the objects. If all the paired comparisons are performed, and if the obtained preference graph has no circuit path, the preference graph represents a complete order between the objects. The occurrence of a circuit path in the preference graph means that certain objects have been cyclically ranked, i.e. in a way that is not consistent with any order.

1.3 Ranking procedures

The reason why a preference graph may contain circuit paths, and consequently not be consistent with any order between the objects, may be that no order exists or that an order exists but the preferences are subject to error. It may be desired to use the preference graph to test whether or not there is any order between the objects. If an order exists, it may be of interest to estimate that order by a ranking based on the preference graph.

Every complete preference graph contains at least one Hamiltonian path, i.e. a path running through each node exactly once. This property of complete preference graphs was shown by Rédei (1934). Proofs have also been provided by König (1936) and Harary, Norman, and Cartwright (1965). Each Hamiltonian path corresponds to an order where every object is preferred to its nearest successor. However, it is not known whether every object is preferred to all its successors, due to the fact that circuit paths may occur.

In incomplete preference graphs it is possible to find, in an analogous way, at least one order where every object is either preferred to its nearest successor or has not been compared to it.

Ranking according to the Hamiltonian principle is unsatisfactory, since there are usually several Hamiltonian paths in the preference graph. For instance, a Hamiltonian circuit path may exist, and then all the cyclic permutations of any order based on it may also be chosen. An example of a complete preference graph with five nodes is shown in Figure 1. All the cyclic permutations of 1, 4, 3, 5, 2 may be used as orders according to the Hamiltonian ranking principle.

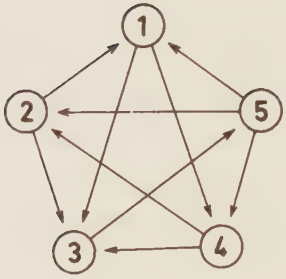


Figure 1

For complete preference graphs Kendall and Smith (1940) have suggested a criterion based on the total number of three-node circuit paths to test whether or not any order exists between the objects. If the existence of an order is accepted, the ranking can be based on the score sequence of the objects. The score of an object is the number of other objects it is preferred to. In graph terminology, the score of an object becomes the local out-degree of the corresponding node. The objects are ranked according to non-increasing score sequence. For the preference graph in Figure 1 the objects 1, 2, 3, 4, 5 have the score sequence 2, 2, 1, 2, 3, and a possible order according to the score ranking principle becomes 5, 1, 2, 4, 3.

Slater (1961) has suggested that the existence or non-existence of an order should be tested by use of a criterion based on the least number of preferences which have to be changed to make the preference graph coincide with a complete order. Each order with the least number of dissentient preferences is called a nearest adjoining order to the preference graph. The preference graph in Figure 1 has 5, 1, 4, 3, 2 as a nearest adjoining order, and there are three preferences which do not agree with that order. No other order has a smaller number of dissentient preferences.

It is considerably more cumbersome to apply Slater's ranking principle than the more common score ranking principle. If the arc frequency matrix of the preference graph is written with the rows and columns in a certain order, the number of preferences which do not agree with that order is obtained as the sum of the matrix elements below the main diagonal. In principle, these numbers

have to be examined for all the N! possible orders of the N objects, and an order corresponding to the least number of dissentient preferences has to be chosen. Slater (1961) suggests a simpler computational algorithm which may be applicable in certain cases. This algorithm is based on cyclic permutations of an order obtained by score ranking. Analogous methods are used by Alway (1962) to show the necessary and sufficient conditions of Landau (1953) that an integer sequence be the score sequence of a preference graph. Another computational algorithm providing the nearest adjoining order by use of dynamic programming has been suggested by Remage and Thompson (1966).

2. STOCHASTIC PREFERENCE GRAPHS

2.1 Preference probabilities

A survey of the stochastic preference models which have been suggested in the literature was given by David (1963). For paired comparisons it is possible to introduce the probability p_{ij} that object i is preferred to object j. Thus, the outcome of the comparison between the objects i and j will be preference for object i with probability p_{ij} and preference for object j with probability $p_{ji} = 1-p_{ij}$.

Ranking by use of the preference probabilities is possible if the following condition of stochastic transitivity holds for each set of three different objects i, j, k:

$$p_{ij} \geq 1/2 \text{ and } p_{jk} \geq 1/2 \text{ implies that } p_{ik} \geq 1/2. \tag{1}$$

This condition is the analog of the fact that a preference graph corresponds to a complete order if, and only if, there is transitivity among the preferences. Another ranking based on the preference probabilities is to rank the objects according to non-increasing mean values of their preference probabilities, i.e. according to non-increasing

$$\sum_j p_{ij}/(N-1) , \tag{2}$$

where the summation is carried out for $j \neq i$.

Ranking based on the preference probabilities presupposes these probabilities to be known. If the preference probabilities are unknown, a basis of the ranking may be provided by observations of the stochastic preference graph. Stochastic problems involved in such ranking procedures will be considered for a particular preference model in the rest of this chapter. First, the preference model will be described and illustrated in the two following subsections, and then the stochastic preference graph of the model will be studied for complete comparison graphs in Section 3 and for complete bipartite comparison graphs in Section 4.

2.2 A preference model

Consider N objects for which a prior order may exist, which can be known or unknown. When a prior order exists, it is assumed that the outcomes of the paired comparisons are independent of each other and in agreement with the prior order with probability p. Thus, all the paired comparisons have the same probability p of resulting in a preference in accordance with the prior order, and probability $q = 1-p$ of resulting in a contrary preference. When no prior order exists, all the preferences are assumed to be given at random, i.e. $p = 1/2$.

Assume that the prior order of the N objects $1, 2, \ldots, N$ is given by the permutation $i_1, i_2, \ldots, i_N$. Then the preference probabilities p_{ij} are equal to p or q, according to whether object i occurs before or after object j in the order $i_1, i_2, \ldots, i_N$. Obviously, stochastic transitivity holds in this case. Moreover, the average preference probability of object i becomes

$$\sum_j p_{ij}/(N-1) = [Np-q-k(p-q)]/(N-1) \qquad (3)$$

if $i = i_k$, i.e. if object i is the k-th object according to the prior order. When $p > 1/2$, the average value (3) is a decreasing function of k, and consequently ranking according to decreasing average preference probabilities will produce the prior order.

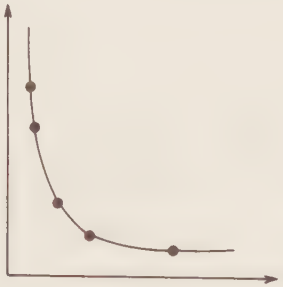


Figure 2

Some problems of interest for this model may be to estimate p when the prior order is unknown, to test fixed values of p, and to estimate the prior order by use of an observed preference graph.

2.3 Some illustrations of the preference model

Three different occasions where the stochastic preference model above is applicable will be briefly discussed.

Assume that the prior order between N objects is known by some of the judges and unknown by the others. Let P denote the proportion of judges that know the prior order. Suppose that the paired comparison experiment is performed in such a way that each comparison is considered by one of the judges chosen at random. The judges are providing preferences in accordance with the prior order if they know it, otherwise they guess. In this way, the preference will be given in accordance with the prior order with probability $p = P + (1-P)/2 = (1+P)/2$. Moreover, the various preferences will be stochastically independent if the judges are independently selected. Thus, the preference model of the preceding subsection is obtained.

Another situation is the following. Suppose that N objects are represented by N points on a decreasing curve, as is exemplified in Figure 2. If the objects are ranked according to increasing coordinates on the horizontal axis, an order is obtained which is the opposite of the order obtained by ranking according to increasing coordinates on the vertical axis. Suppose that the paired comparisons are performed independently of each other, and that the horizontal and vertical

rankings are applied with probabilities p and q, respectively. This corresponds to the preference model of Subsection 2.2, if the order obtained by horizontal ranking is referred to as the prior order.

Finally consider an experiment with paired comparisons between objects which can be correctly ranked according to the prior order. Due to the effects of some disturbance, lack of concentration, misprint on a registration device, or the like, it may happen that the judge gives a wrong preference. If the errors occur independently and with a constant probability q, the preference model of Subsection 2.2 is obtained.

3. COMPLETE PREFERENCE GRAPHS

3.1 Notations

Consider N objects labeled $1,2,\ldots,N$, and assume that all the $\binom{N}{2}$ pairs of different objects are compared n times. The comparison graph has the node set $\Omega = \{1,2,\ldots,N\}$ and the symmetric arc frequency matrix C with elements C_{ij} which are equal to 0 or n as $i = j$ or $i \neq j$ for $i \in \Omega$ and $j \in \Omega$. The number of comparisons where object i is considered is given by the local degree

$$a_i = \sum_j C_{ij} = (N-1)n \tag{4}$$

of node i. The total number of comparisons is given by the arc frequency

$$R = \sum_{i<j} \sum C_{ij} = \binom{N}{2} n \ . \tag{5}$$

Among the n comparisons between object i and object j, the number of preferences for object i is denoted Z_{ij}, and the number of preferences for object j is $Z_{ji} = n - Z_{ij}$. Put $Z_{ii} = 0$ for $i \in \Omega$. The preference graph has the node set Ω and the arc frequency matrix Z with elements Z_{ij} for $i \in \Omega$ and $j \in \Omega$. The number of preferences for object i, i.e. the score of object i, is given by the local out-degree

$$x_i = \sum_j Z_{ij} \tag{6}$$

of node i. Let x denote the score sequence vector. The total number of preferences is equal to the total number of comparisons, i.e.

$$\sum_i x_i = R . \tag{7}$$

The number of preferences that are consistent with a fixed order $i_1, i_2, \ldots, i_N$ of the N objects is denoted $r(i_1, i_2, \ldots, i_N)$. This number can be written

$$r(i_1, i_2, \ldots, i_N) = \sum\sum Z_{ij} , \tag{8}$$

where the summation is carried out for all the pairs (i,j) with i before j in the order $i_1, i_2, \ldots, i_N$. Particularly,

$$r(1,2,\ldots,N) = \sum_{i<j}\sum Z_{ij} . \tag{9}$$

It is possible to interpret $r(i_1, i_2, \ldots, i_N)$ as the arc frequency in the partial graph of the preference graph which is consistent with the order $i_1, i_2, \ldots, i_N$.

3.2 The scores

The probability distribution of the score x_i of object i depends on the prior order. If the prior order is $i_1, i_2, \ldots, i_N$, and if $i = i_k$, then x_i is a sum of N-1 independent binomial variables, of which k-1 are binomial (n,q) and N-k are binomial (n,p). Consequently, x_i is a sum of a binomial ((k-1)n,q) and a binomial ((N-k)n,p) variable, and these variables are independent. Hence the expected value and the variance become

$$E\, x_i = (Np-q)n - (p-q)nk , \tag{10}$$

$$\mathrm{Var}\, x_i = (N-1)npq . \tag{11}$$

When $p > 1/2$, the expected scores are decreasing functions of k, i.e. the expected scores are consistent with the prior order. All the scores have the same variance. The distribution of the relative score x_i/n is asymtotically normal, with expected value $Np-q-(p-q)k$ and variance $(N-1)pq/n$ when n tends to infinity.

The score sequence has the mean value $m(x) = (N-1)n/2$ and the variance

$$s^2(x) = \sum_{i=1}^{N} x_i^2/N - m^2(x) \quad . \tag{12}$$

According to (10), (11), and (12), the expected value of $s^2(x)$ becomes

$$E\, s^2(x) = \sum_{k=1}^{N} [(N-1)npq+(Np-q)^2n^2+(p-q)^2n^2k^2-2(Np-q)(p-q)n^2k]/N -$$

$$- (N-1)^2 n^2/4 \quad . \tag{13}$$

By using the relationships

$$\sum_{k=1}^{N} k/N = (N+1)/2 \tag{14}$$

and

$$\sum_{k=1}^{N} k^2/N = (N+1)(2N+1)/6 \tag{15}$$

in (13), it follows after simplification that

$$E\, s^2(x) = (N-1)npq + (N^2-1)n^2(1-4pq)/12 \quad . \tag{16}$$

For $p = 1/2$, it holds that $E\, s^2(x) = (N-1)n/4$, and simple estimation shows that this value is smaller than the value in (16) when $p \neq 1/2$, $n \geq 1$, and $N > 2$. Consequently, large values of $s^2(x)$ are critical for the hypothesis $p = 1/2$.

Durbin (1951) has studied the distribution of $s^2(x)$ when $p = 1/2$, and he suggests that the distribution of $4s^2(x)/n$ can be approximated by a χ^2_{N-1} distri-

bution when n is large. Starks (1958) has compared this approximation to the exact distribution and found the χ^2_{N-1} approximation reasonably satisfactory in the region of the upper ten per cent point and beyond. The exact distribution of $s^2(x)$ when $p = 1/2$ can be constructed for small values of n and N by use of tables by Trawinski (1961). Some critical values for the sum of squares of scores are given by David (1963) for the combinations $N = 3$ and $n \leq 20$, $N = 4$ and $n \leq 7$, $N = 5$ and $n \leq 3$, $N = 6, 7, 8$ and $n = 1$.

3.3 Agreement

Consider n judges which give their preferences in all the paired comparisons between N objects. Associate with each pair of judges the number of preferences which agree for the two judges. If these numbers are added for all the $\binom{n}{2}$ pairs of judges, a measure h of the agreement between the preferences of the judges is obtained. Another way of obtaining the quantity h is as follows. First it is determined how many pairs of equal preferences there are among the n comparisons between a fixed pair of objects. Then these numbers are added for all the $\binom{N}{2}$ pairs of objects. In this way the measure of agreement is obtained, and it holds that

$$h = \sum_{i<j} \sum \left[\binom{Z_{ij}}{2} + \binom{Z_{ji}}{2} \right] = \binom{N}{2}\binom{n}{2} - \sum_{i<j} \sum Z_{ij} Z_{ji} . \qquad (17)$$

If for each pair of objects their n preferences agree, h will take its maximum value $\binom{N}{2}\binom{n}{2}$. The minimum value of h is obtained if for each pair of objects there are n/2 preferences of each kind when n is even and (n-1)/2 preferences of one kind and (n+1)/2 preferences of the other kind when n is odd. Thus, the minimum value of h becomes $\binom{N}{2}n(n-2)/4$ when n is even and $\binom{N}{2}(n-1)^2/4$ when n is odd.

The measure h of agreement has been introduced by Kendall and Smith (1940). Kendall (1962) provides tables of the distribution of h when $p = 1/2$ and N and n take on small values. Generalizations to arbitrary p will be studied here. These cases are of interest when for instance it is desired to calculate the power of

tests of the hypothesis $p = 1/2$ against alternative values $p \neq 1/2$.

It is possible to write $Z_{ij}\, Z_{ji} = Z(n-Z)$ where Z is binomial (n,p). In fact, if the prior order has object i before j, then $Z_{ij} = Z$ and otherwise $Z_{ji} = Z$. It follows immediately that

$$E\, Z(n-Z) = n\, E\, Z - E\, Z^2 = n(n-1)pq \quad . \tag{18}$$

In order to determine the variance, it is possible to use the moments of the binomial distribution up to the fourth order, and by lengthy algebraic calculations it follows that

$$\text{Var}\, Z(n-Z) = 2\binom{n}{2}pq(1-2pq) + 6\binom{n}{3}pq(1-4pq) \quad . \tag{19}$$

More easily the variance is obtained if independent Bernoulli (p) variables Z_i are introduced. Put $Z_i' = 1-Z_i$ and

$$Z(n-Z) = \sum_{i=1}^{n} \sum_{j=1}^{n} Z_i\, Z_j' = \sum\sum_{i \neq j} Z_i\, Z_j' \quad . \tag{20}$$

It follows that

$$\text{Var}\, Z(n-Z) = \sum\sum_{i \neq j}\ \sum\sum_{k \neq l} \text{Cov}\,(Z_i\, Z_j',\, Z_k\, Z_l') \quad , \tag{21}$$

and

$$\text{Cov}\,(Z_i Z_j',\, Z_k Z_l') = \begin{cases} pq - p^2q^2 & \text{if } k=i \text{ and } l=j, \\ -p^2q^2 & \text{if } k=j \text{ and } l=i, \\ pq^2 - p^2q^2 & \text{if } k=i \text{ and } l \neq i,j, \\ -p^2q^2 & \text{if } k=j \text{ and } l \neq i,j, \\ -p^2q^2 & \text{if } k \neq i,j \text{ and } l=i, \\ p^2q - p^2q^2 & \text{if } k \neq i,j, \text{ and } l=j, \\ 0 & \text{if } k \neq i,j \text{ and } l \neq i,j, \end{cases} \tag{22}$$

which implies that (19) holds.

From (17) and (18) it follows that

$$E\,h = \binom{N}{2}\binom{n}{2}(1-2pq) \quad , \tag{23}$$

and since the products $Z_{ij}\,Z_{ji}$ are independent for $i < j$, it follows from (17) and (19) that

$$\mathrm{Var}\,h = 2\binom{N}{2}\binom{n}{2}pq(1-2pq) + 6\binom{N}{2}\binom{n}{3}pq(1-4pq) \quad . \tag{24}$$

Particularly, for $p = 1/2$

$$E\,h = \binom{N}{2}\binom{n}{2}/2 \quad , \tag{25}$$

and

$$\mathrm{Var}\,h = \binom{N}{2}\binom{n}{2}/4 \quad . \tag{26}$$

According to the central limit theorem, h is asymptotically normal when N tends to infinity. Another approximation is obtained when n tends to infinity, if the distribution of Z is approximated by a normal distribution. Put $Y = (Z-np)/\sqrt{npq}$ and use the identity

$$Z(n-Z) = n^2/4 - npq\left[Y+n(p-q)/2\sqrt{npq}\right]^2 \quad . \tag{27}$$

The formula (17) can then be rewritten

$$h = \binom{N}{2}n(n-2)/4 + npq \sum_{i<j}\sum \left[Y_{ij} + \sqrt{(1-4pq)n/4pq}\right]^2 \quad , \tag{28}$$

with independent approximate standard normal variables Y_{ij}. It follows that, when n tends to infinity,

$$\left[h - \binom{N}{2}n(n-2)/4\right]/npq \tag{29}$$

is approximately non-central $\chi^2_\nu(\delta)$-distributed with $\nu = \binom{N}{2}$ degrees of free-

dom and non-centrality parameter

$$\delta = \sqrt{\binom{N}{2}(1-4pq)n/4pq} \tag{30}$$

according to the terminology used by Scheffé (1959). Patnaik (1949) has suggested that non-central $\chi^2_v(\delta)$ be approximated by $c\,\chi^2_u$, where the constant c and the degrees of freedom u are determined by equating the means and variances of the two distributions. This gives $c = (v+2\delta^2)/(v+\delta^2)$ and $u = (v+\delta^2)^2/(v+2\delta^2)$. Some numerical evidence of this approximation to non-central χ^2 is given by Tukey (1957). In the particular case $p = 1/2$ it follows that $4h/n-\binom{N}{2}(n-2)$ is approximately χ^2-distributed with $\binom{N}{2}$ degrees of freedom.

A third approximation when $p \neq 1/2$ is obtained if (20) is rewritten

$$Z(n-Z) = n \sum_{i=1}^{n} (Z_i-\bar{Z})^2 , \tag{31}$$

with $\bar{Z} = \Sigma Z_i/n$. According to Cramér (1957, § 28.3), the variable $Z(n-Z)/n^2$ for increasing n is asymptotically normal, with expected value equal to $\mathrm{Var}\, Z_i = pq$ and variance equal to

$$[E(Z_i-p)^4 - p^2q^2]/n = pq(1-4pq)/n . \tag{32}$$

Application of these results to the independent variables $Z_{ij}\,Z_{ji}/n^2$ gives that h/n^2 for increasing n has an asymptotically normal distribution, with expected value $\binom{N}{2}(1-2pq)/2$ and variance $\binom{N}{2}pq(1-4pq)/n$.

3.4 Inconsistency

Kendall and Smith (1940) have introduced the total number t of three-node circuit paths as a measure of the inconsistency of the preferences in a complete preference graph. They show that for $n = 1$, the lowest possible value of t is zero, and the highest possible value of t is $N(N^2-1)/24$ if N is odd and $N(N^2-4)/24$ if N is even. It follows that among the $\binom{N}{3}$ subgraphs with three

nodes, the proportion of circuits tends to 1/4 when N tends to infinity. It is interesting that the stochastic preference graphs with $n = 1$ and $p = 1/2$ have $E\,t = \binom{N}{3}/4$, i.e. the expected proportion of circuits is exactly 1/4. Moran (1947) has shown that $\mathrm{Var}\,t = 3\binom{N}{3}/16$ when $n = 1$ and $p = 1/2$. The case of an arbitrary p has been considered by Frank (1968), who derives formulas for $E\,t$ and $\mathrm{Var}\,t$ when $n = 1$. These results will here be generalized to arbitrary n.

The number t of circuits can be written

$$t = \sum_{i<j<k} \sum \sum (Z_{ij}Z_{jk}Z_{ki} + Z_{kj}Z_{ji}Z_{ik}) = \sum_u t_u \tag{33}$$

as a sum of $\binom{N}{3}$ variables t_u associated with the different three-node subgraphs. Whatever the prior order may be, each variable t_u can be represented by

$$t_u = Z_1Z_2Z_3' + Z_1'Z_2'Z_3 \ , \tag{34}$$

with Z_i independent binomial (n,p) and $Z_i' = n - Z_i$. It follows that $E\,t_u = n^3pq$, and consequently

$$E\,t = \binom{N}{3}n^3pq \ . \tag{35}$$

Moreover,

$$\mathrm{Var}\,t = \sum_u \sum_v \mathrm{Cov}\,(t_u, t_v) \ , \tag{36}$$

where the sum contains $\binom{N}{3}^2$ terms, of which $\binom{N}{3}$ refer to three nodes, $\binom{N}{3}\binom{3}{2}\binom{N-3}{1} = 12\binom{N}{4}$ refer to four nodes, $\binom{N}{3}\binom{3}{1}\binom{N-3}{2} = 30\binom{N}{5}$ refer to five nodes, and $\binom{N}{3}\binom{N-3}{3} = 20\binom{N}{6}$ refer to six nodes. The variables t_u and t_v are independent if they are associated with subgraphs without common arcs. This implies that the terms of (36) which refer to five or six nodes are equal to zero. It remains to determine $\mathrm{Var}\,t_u$ and $\mathrm{Cov}\,(t_u, t_v)$ when t_u and t_v are associated with subgraphs having two nodes in common. By use of the representation (34) and the relationships

$$E\, Z_i^2 = npq + n^2p^2 \quad , \tag{37}$$

$$E\, (Z_i')^2 = npq + n^2q^2 \quad , \tag{38}$$

$$E\, Z_i Z_i' = n(n-1)pq \quad , \tag{39}$$

it follows that

$$\mathrm{Var}\, t_u = 3n^4p^2q^2 + n^5pq(1-4pq) \quad . \tag{40}$$

Consider then t_u and t_v associated with two subgraphs with two nodes in common. Let the four nodes that are involved be i,j,k,l, taken in the same order as they appear in the prior order. There are twelve different terms in the sum (36) which refer to exactly these four nodes. Since $\mathrm{Cov}\,(t_u,t_v) = \mathrm{Cov}\,(t_v,t_u)$, only six different terms have to be considered. The six cases correspond to the six possibilities of choosing the two common nodes for the two subgraphs. If the common

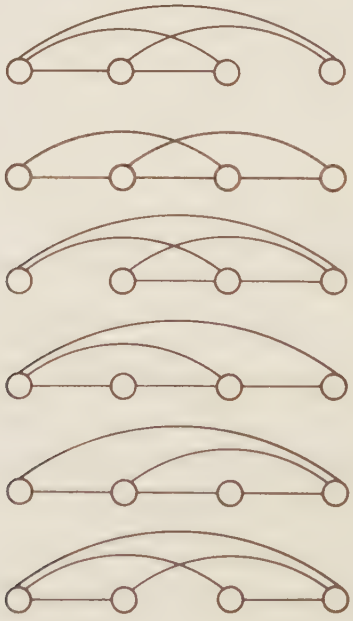


Figure 3

nodes are ij, jk, kl, ik, jl, il, respectively, the two subgraphs contain the respective nodes

1. ijk and ijl ,
2. ijk and jkl ,
3. ikl and jkl ,
4. ijk and ikl ,
5. ijl and jkl ,
6. ijl and ikl .

Put for the sake of brevity, $Z_1 = Z_{ij}$, $Z_2 = Z_{jk}$, $Z_3 = Z_{kl}$, $Z_4 = Z_{ik}$, $Z_5 = Z_{jl}$, and $Z_6 = Z_{il}$. Then the product $t_u t_v$ can be written in the following way in the six cases

$$
\left.
\begin{aligned}
&1.\quad (Z_1 Z_2 Z_4' + Z_1' Z_2' Z_4)(Z_1 Z_5 Z_6' + Z_1' Z_5' Z_6) ,\\
&2.\quad (Z_1 Z_2 Z_4' + Z_1' Z_2' Z_4)(Z_2 Z_3 Z_5' + Z_2' Z_3' Z_5) ,\\
&3.\quad (Z_3 Z_4 Z_6' + Z_3' Z_4' Z_6)(Z_2 Z_3 Z_5' + Z_2' Z_3' Z_5) ,\\
&4.\quad (Z_1 Z_2 Z_4' + Z_1' Z_2' Z_4)(Z_3 Z_4 Z_6' + Z_3' Z_4' Z_6) ,\\
&5.\quad (Z_1 Z_5 Z_6' + Z_1' Z_5' Z_6)(Z_2 Z_3 Z_5' + Z_2' Z_3' Z_5) ,\\
&6.\quad (Z_1 Z_5 Z_6' + Z_1' Z_5' Z_6)(Z_3 Z_4 Z_6' + Z_3' Z_4' Z_6) .
\end{aligned}
\right\} \qquad (41)
$$

The six cases are shown in Figure 3, where the four nodes are placed from left to right according to the prior order. The variables t_u and t_v have the common arc variable Z_k in the case labeled k. An examination of (41) shows that the six cases reduce to only three essentially different cases. In fact, the cases 1, 2, and 3 contain the common arc variable in analogous ways. Also, the cases 4 and 5 are analogous. Consequently it suffices to consider cases 1, 4, and 6. It may be noted that the discriminating characteristic of the six cases is the number of nodes occurring between the two common nodes according to the prior order. In cases 1, 2, and 3 there is no node in between, in cases 4 and 5 there is one node in between, and in case 6 there are two nodes in between.

By straightforward calculations, using (37), (38), and (39), it is found that

$$\text{Cov}\,(t_u, t_v) = 0 \text{ in cases 1, 2, 3, 4, 5,} \tag{42}$$

and

$$\text{Cov}\,(t_u, t_v) = n^5 pq(1-4pq) \text{ in case 6.} \tag{43}$$

Thus we have the unexpected result that t_u and t_v are uncorrelated in all cases except when their node triplets have the two nodes in common which are the best and worst nodes among the four nodes involved. Here, best and worst refer to the prior order.

As a consequence of the above results, the sum (36) contains $\binom{N}{3}$ terms which are equal to the expression (40) and $2\binom{N}{4}$ terms which are equal to the expression (43). Thus it holds that

$$\text{Var } t = 3\binom{N}{3} n^4 p^2 q^2 + \binom{N}{3}(N-1) n^5 pq(1-4pq)/2 \ . \tag{44}$$

3.5 Relationship between agreement and inconsistency

In this subsection it will be shown that there exists a linear relationship between the score variance $s^2(x)$, the agreement measure h, and the inconsistency measure t. This relationship is a generalization of the relationship

$$t = \binom{N+1}{3}/4 - Ns^2(x)/2 \ , \tag{45}$$

which holds for n = 1 and is given by David (1963).

To derive the general relationship for arbitrary n, the formula (33) for t is written

$$t = \sum_{\neq}\sum\sum Z_{ij} Z_{jk} Z_{ki}/3 = \sum_{\neq}\sum\sum (n-Z_{ji})(n-Z_{kj})(n-Z_{ik})/3 \ , \tag{46}$$

where the summations are carried out for all the N(N-1)(N-2) choices of differ-

ent indices i,j,k. By expanding the right-hand side of (46), the following equation is obtained:

$$t = 2\binom{N}{3}n^3 - (N-2)n^2 \sum_{\neq}\sum Z_{ij} + n \sum_{\neq}\sum\sum Z_{ij}Z_{jk} - t \ . \tag{47}$$

Here,

$$\sum_{\neq}\sum Z_{ij} = \binom{N}{2}n \ , \tag{48}$$

and according to (12) and (17)

$$\sum_{\neq}\sum\sum Z_{ij}Z_{jk} = \sum_{\neq}\sum Z_{ij}(x_j - Z_{ji}) = \binom{N}{2}(N-1)n^2 - \sum x_j^2 - \sum_{\neq}\sum Z_{ij}Z_{ji} =$$

$$= 2h - Ns^2(x) + \binom{N}{2}(N-1)n^2/2 - 2\binom{N}{2}\binom{n}{2} \ . \tag{49}$$

Substitution of (48) and (49) in (47) gives the relationship

$$t = \binom{N+1}{3}n^3/4 - \binom{N}{2}\binom{n}{2}n + nh - Nns^2(x)/2 \ . \tag{50}$$

3.6 Complete orders

When a complete stochastic preference graph is free from circuit paths, i.e. when t = 0, the graph represents a complete order which may or may not coincide with the prior order. The probability that all the preferences are consistent with an order will be determined. This probability P(t = 0) can be obtained as the sum of the probabilities associated with the N! preference graphs which have all their $R = \binom{N}{2}n$ preferences consistent with an order. A preference graph representing the complete order $i_1, i_2, \ldots, i_N$ has $Z_{ij} = n$ for all the $\binom{N}{2}$ pairs (i,j) where i appears before j in the order $i_1, i_2, \ldots, i_N$. If exactly k of these pairs have i before j according to the prior order, the preference graph has the probability

$$p^{nk} q^{R-nk} . \tag{51}$$

If $a_N(k)$ denotes the number of orders $i_1, i_2, \ldots, i_N$ which have exactly k preferences in accordance with the prior order, it holds that

$$P(t=0) = \sum_{k=0}^{\binom{N}{2}} a_N(k) p^{nk} q^{R-nk} . \tag{52}$$

It is no restriction to assume that the objects are labeled according to the prior order, i.e. that the prior order is $1, 2, \ldots, N$. Then it is possible to interpret $a_N(k)$ as the number of permutations $i_1, i_2, \ldots, i_N$ of $1, 2, \ldots, N$ which have exactly $\binom{N}{2}$-k inversions. By symmetry, $a_N(k) = a_N(\binom{N}{2}-k)$ for $k = 0, 1, \ldots, \binom{N}{2}$. Define $a_N(k) = 0$ for $k < 0$. The following relationship holds between $a_N(k)$ and $a_{N-1}(k)$, $a_{N-1}(k-1), \ldots, a_{N-1}(k-N+1)$. Among the $a_N(k)$ permutations of $1, 2, \ldots, N$ which have k inversions, there are $a_{N-1}(k)$ which have object N in the last position, $a_{N-1}(k-1)$ which have object N in the last but one position, etc. This is a consequence of the fact that object N gives inversions with all the objects coming afterwards. Thus it holds that

$$a_N(k) = \sum_{j=0}^{N-1} a_{N-1}(k-j) \tag{53}$$

for $N = 2, 3, \ldots$, with the initial value $a_1(0) = 1$. Define the generating function

$$A_N(u) = \sum_{k=0}^{\binom{N}{2}} a_N(k) u^k . \tag{54}$$

From (53) it follows that

$$A_N(u) = \sum_{j=0}^{N-1} u^j A_{N-1}(u) = \begin{cases} (u^N-1)A_{N-1}(u)/(u-1) & \text{if } u \neq 1 , \\ N A_{N-1}(1) & \text{if } u = 1. \end{cases} \tag{55}$$

Since $A_1(u) = 1$, it follows from (55) by iterations that

$$A_N(u) = \begin{cases} \prod_{k=1}^{N} (u^k-1)/(u-1)^N & \text{if } u \neq 1 , \\ N! & \text{if } u = 1 . \end{cases} \tag{56}$$

The desired probability (52) can be written

$$P(t=0) = q^R A_N((p/q)^n) , \tag{57}$$

and it follows from (56) that

$$P(t=0) = \begin{cases} \prod_{k=1}^{N} (p^{nk}-q^{nk})/(p^n-q^n)^N & \text{if } p \neq 1/2 , \\ N!/2^R & \text{if } p = 1/2 . \end{cases} \tag{58}$$

This probability has been obtained by Frank (1968) for the case $n = 1$. The case $n = 1$ and $p = 1/2$ has been considered by Moran (1947), who shows that t has an asymptotically normal distribution when N tends to infinity. Kadane (1966) gives an algorithmic method of calculating $P(t=k)$ for arbitrary k in the case $n = 1$ and $p = 1/2$.

3.7 Estimation of the preference probability

The likelihood function of the preference graph becomes

$$\prod \binom{n}{Z_{ij}} p^{Z_{ij}} q^{n-Z_{ij}} , \tag{59}$$

where the product is carried out for all the $\binom{N}{2}$ pairs (i,j) where i comes before j in the prior order $i_1, i_2, \ldots, i_N$. Since

$$\binom{n}{Z_{ij}} = \binom{n}{n-Z_{ij}} = \binom{n}{Z_{ji}} , \tag{60}$$

the likelihood function attains its maximum value for the same prior order $i_1, i_2, \ldots, i_N$ and the same preference probability p as the function

$$r \log p + (R-r) \log q . \tag{61}$$

Here $r = r(i_1, i_2, \ldots, i_N)$ denotes the number of preferences which are consistent with the prior order $i_1, i_2, \ldots, i_N$. When the prior order is known, the maximum likelihood estimator of p becomes equal to $p^* = r/R$, and r is binomial (R,p). When the prior order is unknown, the function (61) has to be maximized both with respect to permutations $i_1, i_2, \ldots, i_N$ and probabilities p. Then it follows by a standard argument that p has the maximum likelihood estimator $p^* = r/R$, where $r = r(i_1, i_2, \ldots, i_N)$ is given by an order $(i_1, i_2, \ldots, i_N)$ which maximizes

$$p^* \log p^* + (1-p^*) \log (1-p^*) . \tag{62}$$

This expression is the negative of the entropy in the terminology of information theory (Kullback, 1959). It may be said that the prior order is estimated by an order of minimum entropy, and the preference probability is estimated by the relative frequency of preferences which are consistent with that order. Thompson and Remage (1964) have obtained an analogous result for another preference model. It may be noted that the estimator p^* is not unique; it is impossible to separate p^* and $1-p^*$. If $i_1, i_2, \ldots, i_N$ is an order which maximizes (62), the same is true of the reversed order $i_N, i_{N-1}, \ldots, i_1$. Since (62) has maximum at the same time as $|p^*-1/2|$, an estimator of the prior order can simply be characterized as a nearest adjoining order, i.e. an order that maximizes $r(i_1, i_2, \ldots, i_N)$.

The difficulties involved in the determination of a nearest adjoining order (Slater, 1961; Remage and Thompson, 1966) may be reason to look for other estimators of p. Since p and q are indistinguishable, it seems convenient to estimate the product $\theta = pq$. If it is known that $p > 1/2$, it easily follows that $p = (1 + \sqrt{1 - 4\theta})/2$. To estimate θ, the agreement measure h or the inconsistency measure t can be used. According to (23) and (35) it follows that θ has the unbiased estimators

$$\theta^* = 1/2 - h/2\binom{N}{2}\binom{n}{2} , \tag{63}$$

and

$$\hat{\theta} = t/\binom{N}{3}n^3 . \tag{64}$$

From the variance formulas (24) and (44) it follows that

$$\text{Var } \theta^* = \theta(1-2\theta)/2\binom{N}{2}\binom{n}{2} + 3\binom{n}{3}\theta(1-4\theta)/2\binom{N}{2}\binom{n}{2}^2 , \tag{65}$$

and

$$\text{Var } \hat{\theta} = 3\theta^2/\binom{N}{3}n^2 + (N-1)\theta(1-4\theta)/2\binom{N}{3}n . \tag{66}$$

Hence it follows that

$$n(\text{Var } \theta^* - \text{Var } \hat{\theta}) \to (3N-7)\theta(1-4\theta)/2\binom{N}{3} , \tag{67}$$

when n tends to infinity. Since the limit value is positive for $N > 2$, $\hat{\theta}$ is asymptotically more efficient that θ^* .

4. COMPLETE BIPARTITE PREFERENCE GRAPHS

4.1 Introduction

Suppose that the N objects which are to be compared are separated into two sets Ω_1 and Ω_2 with N_1 and $N_2 = N-N_1$ objects, respectively. All the paired comparisons between the sets are performed, but no comparisons within the sets. If n comparisons are performed for each pair (i,j) of objects $i \in \Omega_1$ and $j \in \Omega_2$, a total of $R = N_1 N_2$ n comparisons are obtained. The comparison graph becomes a bipartite graph with N_1 and N_2 nodes in its two node subsets and R arcs in between. The n comparisons between object $i \in \Omega_1$ and object $j \in \Omega_2$ results in Z_{ij} preferences for object i and $Z_{ji} = n-Z_{ij}$ preferences for object j. Put $Z_{ij} = 0$ if $i \in \Omega_1$ and $j \in \Omega_1$, or if $i \in \Omega_2$ and $j \in \Omega_2$. The score of object $i \in \Omega_1$ becomes

$$x_i = \sum_{j \in \Omega_2} Z_{ij} , \tag{68}$$

and the score of object $j \in \Omega_2$ becomes

$$x_j = \sum_{i \in \Omega_1} Z_{ji} = N_1 n - \sum_{i \in \Omega_1} Z_{ij} . \tag{69}$$

The number $r(i_1, i_2, \ldots, i_N)$ of preferences which are consistent with the order $i_1, i_2, \ldots, i_N$ can be written

$$r = \sum_{i \in \Omega_1} \sum_{j \in \Omega_2} Z_{ij} \tag{70}$$

for each order having the property that all the Ω_1 objects come before all the Ω_2 objects. For this kind of prior order, Z_{ij} will be binomial (n,p) and independent for $i \in \Omega_1$ and $j \in \Omega_2$. Then it follows easily, that the scores x_i are binomial $(N_2 n, p)$ and independent for $i \in \Omega_1$, the scores x_j are binomial $(N_1 n, q)$ and independent for $j \in \Omega_2$, and r in (70) is binomial (R,p). More interesting than this simple case are such prior orders where the Ω_1 and Ω_2 objects are mixed up.

4.2 Agreement

The treatment of the agreement measure h in Subsection 3.3 can be carried through in an analogous way for graphs more general than the complete graphs. For complete bipartite graphs it holds that

$$h = N_1 N_2 \binom{n}{2} - \sum_{i \in \Omega_1} \sum_{j \in \Omega_2} Z_{ij} Z_{ji} , \tag{71}$$

and according to (18),

$$E\, h = N_1 N_2 \binom{n}{2}(1-2pq) , \tag{72}$$

and according to (19),

$$\mathrm{Var}\, h = N_1 N_2 \left[2\binom{n}{2}pq(1-2pq) + 6\binom{n}{3}pq(1-4pq)\right] . \tag{73}$$

Moreover, the variable

$$[h - N_1 N_2 n(n-2)/4]/npq \tag{74}$$

is approximately non-central $\chi^2_\nu(\delta)$-distributed with $\nu = N_1 N_2$ degrees of freedom and non-centrality parameter

$$\delta = \sqrt{N_1 N_2\, n(1-4pq)/4pq} \tag{75}$$

when n tends to infinity. For $p = 1/2$, this becomes an ordinary χ^2-distribution. For $p \neq 1/2$, it also holds that h/n^2 is asymptotically normal, with mean value $N_1 N_2(1-2pq)/2$ and variance $N_1 N_2 pq(1-4pq)/n$ when n tends to infinity.

4.3 Inconsistency

For complete graphs, the existence of a circuit path implies the existence of a three-node circuit path. Analogously, for complete bipartite graphs, the existence of a circuit path implies the existence of a four-node circuit path.

Each four-node circuit path contains two Ω_1 nodes and two Ω_2 nodes. The total number t of four-node circuit paths may be used as a measure of the inconsistency among the preferences. It is possible to write $t = \sum_u t_u$ as a sum of $\binom{N_1}{2}\binom{N_2}{2}$ variables t_u associated with the different four-node subgraphs containing two Ω_1 nodes and two Ω_2 nodes.

Consider a four-node subgraph of the preference graph. One has to distinguish between some different cases depending on the positions of the Ω_1 and Ω_2 nodes in the prior order. In Figure 4, three cases are shown which are characterized as follows.

1. Ω_1 and Ω_2 nodes alternate.
2. Ω_1 (Ω_2) nodes are situated in the two middle positions.
3. Ω_1 (Ω_2) nodes are situated in the first two positions.

The prior order is from left to right in Figure 4, and the Ω_1 and Ω_2 nodes are distinguished by different marks. In the first case, it is possible to write

$$t_u = Z_1 Z_2 Z_3 Z_4' + Z_1' Z_2' Z_3' Z_4 \ , \tag{76}$$

with Z_i independent binomial (n,p) and $Z_i' = n - Z_i$ for i = 1,2,3,4. In the second and third cases,

$$t_u = Z_1 Z_2 Z_3' Z_4' + Z_1' Z_2' Z_3 Z_4 \ . \tag{77}$$

Hence

$$E\, t_u = \begin{cases} n^4 pq(1-2pq) \text{ in case 1,} \\ 2n^4 p^2 q^2 \text{ in cases 2 and 3.} \end{cases} \tag{78}$$

To obtain a formula of E t, one has to introduce a parameter m which gives the number of four-node subgraphs in which the Ω_1 and the Ω_2 nodes are alternat-

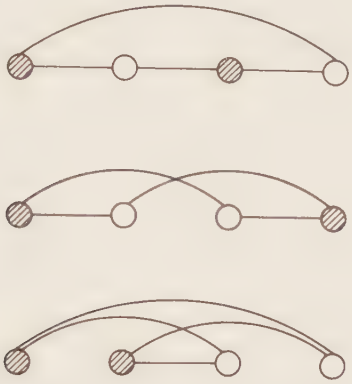


Figure 4

ing when the four nodes are ordered according to the prior order. The parameter m gives a number that characterizes the prior order. It may be noted that m equals zero if, and only if, all the Ω_1 nodes are situated before or after all the Ω_2 nodes in the prior order. It follows from (78) that

$$E\, t = m\, n^4 pq(1-4pq) + \binom{N_1}{2}\binom{N_2}{2} 2\, n^4 p^2 q^2 \ . \tag{79}$$

The variance of t can be determined by use of the covariance of the variables t_u. Only the simplest case, $p = 1/2$, will be considered here. Then it follows by straightforward calculations that

$$\mathrm{Var}\ t_u = (n^4 + 6n^6)/64 \ . \tag{80}$$

When t_u and t_v are associated with two subgraphs without common Ω_1 nodes or without common Ω_2 nodes, they have no common arc variables, and consequently they are then independent. When t_u and t_v are associated with two subgraphs with one common Ω_1 node and one common Ω_2 node, they have one common arc variable, and it follows by straightforward calculations that they are uncorrelated. When t_u and t_v are associated with two subgraphs with three common nodes, they have two common arc variables, and it follows by straightforward calculations that

$$\text{Cov}\,(t_u, t_v) = n^6/64 \quad . \tag{81}$$

Then by a standard combinatorial argument it follows from (80) and (81) that

$$\text{Var}\, t = \sum_u \sum_v \text{Cov}\,(t_u, t_v) = \binom{N_1}{2}\binom{N_2}{2}\,(n^4 + 6n^6)/64 +$$

$$+ \binom{N_1}{2}\binom{N_2}{2}\, 2(N_1+N_2-4)n^6/64 =$$

$$= \binom{N_1}{2}\binom{N_2}{2}\left[n^4 + 2n^6(N_1+N_2-1)\right]/64 \quad . \tag{82}$$

REFERENCES

Alway, G.G., 1962. Matrices and sequences. Math.Gaz. 46, 208-213.

Austin, T.L., Fagen, R.E., Penney, W.F., Riordan, J., 1959. The number of components in random linear graphs. Ann.Math.Statist. 30, 747-754.

Avondo-Bodino, G., 1962. Economic applications of the theory of graphs. Gordon and Breach, New York.

Barton, D.E., David, F.N., 1966. The random intersection of two graphs. In David, F.N., (Ed.): Festschrift for J. Neyman. Research papers in statistics. Wiley, New York.

Bavelas, A., 1950. Communication patterns in task-oriented groups. J.Acoust. Soc.Amer. 57, 271-282.

Beauchamp, C., 1965. Un meilleur indice de centralité. Behavioral Sci. 11, 161-163.

Ben-Israel, A., 1965. An iterative method of computing the generalized inverse of an arbitrary matrix. Math.Comput. 10, 452-455.

Berge, C., Ghouila-Houri, A., 1965. Programming, games and transportation networks. Methuen, London.

Berman, E.B., 1964. Resource allocation in a PERT network under continuous time-cost functions. Management Sci. 10, 734-745.

Bhargava, T.N., 1965. On time changes in a digraph. J.Appl.Probability 2, 101-118.

Bhargava, T.N., Fisk, D.L., 1967. Random graphs and probability models. Paper presented at the 36th session of the International Statistical Institute, Sydney, Australia.

Blackwell, W.A., 1968. Mathematical modeling of physical networks. Mac Millan, New York.

Bloemena, A.R., 1964. Sampling from a graph. Math.Centre Tracts, Amsterdam.

Boot, J.C.G., 1963. The computation of the generalized inverse of singular or rectangular matrices. Amer.Math. Monthly, 302-303.

Borenius, G., 1969. Note on a problem of O. Frank concerning flow estimation. FOA P rapport C 8230-13. Research Institute of National Defence, Stockholm.

Bronfenbrenner, U., 1943. A constant frame of reference for sociometric research. Sociometry 6. (Reprinted in Sociometry monograph No. 6, Beacon House, New York, 1945.)

Brown, D.T., 1959. A note on approximations to discrete probability distributions. Information and Control 2, 386-392.

Buckner, H.T., 1965. A theory of rumor transmissions. Public Opinion Quart. 29, 54.

Chomsky, N., 1968. Syntactic structures, Mouton, Haag.

Cochrun, B.L., 1967. Transistor circuit engineering. Mac Millan, New York.

Coleman, J.S., 1958. Relational analysis. The study of social organizations with survey methods. Human Organization 17, 28-36.

Coleman, J.S., Katz, E., Menzel, H., 1957. The diffusion of an innovation among physicians. Sociometry 20, 253.

Cramér, H., 1947. Mathematical methods of statistics. Princeton University Press, Princeton.

Criswell, J.H., 1947. The measurement of group integration. Sociometry 10. (Reprinted in Moreno, J.L. et al. (Ed.): The sociometry reader. Free Press, Glencoe, 1960.)

Danzig, G.B., 1960. On the shortest route through a network. Management Sci. 6, 187-190.

Danzig, G.B., 1963. Linear programming and extensions. Princeton University Press, Princeton.

David, H.A., 1963. The method of paired comparisons. Griffin, London.

Deming, W.E., Stephan, F.F., 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. Ann.Math.Statist. 11, 427-444.

Durbin, J., 1951. Incomplete blocks in ranking experiments. Brit.J.Psychol. 4, 85-90.

Edwards, D.S., 1948. The constant frame of reference problem in sociometry. Sociometry 11, 372-379.

El-Badry, M.A., Stephan, F.F., 1955. On adjusting sample tabulations to census counts. J.Amer.Statist.Assoc. 50, 738-762.

Erdös, P., Rényi, A., 1959. On random graphs. Publ.Math. Debrecen 6, 290-297.

Erdös, P., Rényi, A., 1960. On the evolution of random graphs. Publ.Math. Inst. Hungar.Acad.Sci. 5, 17-61.

Feller, W., 1957. An introduction to probability theory and its applications. Vol. 1. Wiley, New York.

Festinger, L., 1949. The analysis of sociograms using matrix algebra. Human Relations 2, 153-158. (Reprinted in Moreno, J.L. et al. (Ed.): The sociometry reader. Free Press, Glencoe, 1960.)

Fienberg, S.E., 1970. An iterative procedure for estimation in contingency tables. Ann.Math.Statist. 41, 907-917.

Flament, C., 1963. Applications of graph theory to group structure. Prentice-Hall, Englewood Cliffs.

Floyd, R.W., 1963. Syntactic analysis and operator precedence. J.Assos. Computing Machinery 10, 3.

Ford, L.R., 1956. Network flow theory. RAND Corp. Paper P-923.

Ford, L.R., Fulkerson, D.R., 1962. Flows in networks. Princeton University Press, Princeton.

Frank, O., 1966. Orienterade och stokastiska grafer. Kompendium No. 19. Research Institute of National Defence, Stockholm.

Frank, O., 1968. Stochastic competition graphs. Rev.Inst.Internat.Statist. 36, 319-326.

Frank, O., 1969 a. Structure inference and stochastic graphs. FOA Reports 3, 2, 1-8.

Frank. O.. 1969 b. Estimation of the flow in tree graphs. FOA Reports 3, 6, 1-10.

Frank, O., 1969 c. Estimation of flows in networks. FOA P rapport C 8237-13. Research Institute of National Defence, Stockholm.

Fréchet. M., 1940, 1943. Les probabilités associées à un système d'événements compatibles et dépendants. Actualités scientifique et industrielles. Paris. Nos. 859 and 942.

Fulkerson, D.R., 1961. A network flow computation for project cost curves. Management Sci. 7, 167-178.

Fulkerson, D.R., 1962. Expected critical path lengths in PERT networks. Operations Res. 10, 808-817.

Gale. D., 1960. The theory of linear economic models. McGraw-Hill, New York.

Gilbert, E.N., 1959. Random graphs. Ann.Math.Statist. 30, 1141-1144.

Gill, A., 1962. Introduction to the theory of finite-state machines. McGraw-Hill, New York.

Ginsburg, S.. 1962. An introduction to mathematical machine theory. Addison-Wesley, Reading.

Goodman, L.A., 1961. Snowball sampling. Ann.Math.Statist. 32, 148-170.

Graybill. F.A., 1969. Introduction to matrices with applications in statistics. Wadsworth, Belmont.

Grossman, I.F., Lerchs, H., 1961. An algorithm for directed graphs with application to the project cost curve and in-process inventory. Paper presented at the Third Annual Conference of the Canadian Operational Research Society, Ottawa.

Guilford, J.P., 1954. Psychometric methods. McGraw-Hill, New York.

Hammer, P.L., 1965. Some network-flow problems solved with pseudo-Boolean programming. Operations Res. 13, 388-399.

Hammer, P.L., Rosenberg, I., 1964. Application of pseudo-Boolean programming to the theory of graphs. Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 3, 167-176.

Hammer, P.L., Rudeanu, S., 1968. Boolean methods in operations research and related areas. Springer, Berlin, Heidelberg.

Harary, F., 1955. The number of linear, directed, rooted and connected graphs. Trans.Amer.Math.Soc. 78, 445-463.

Harary, F., 1967. Graphical enumeration problems. In Harary, F., (Ed.): Graph theory and theoretical physics. Academic Press, New York.

Harary, F., Ross, I.C., 1952. On the determination of redundancies in sociometric chains. Psychometrika 17, 195-208.

Harary, F., Norman, R.Z., Cartwright, D., 1965. Structural models. An introduction to the theory of directed graphs. Wiley, New York.

Hildenbrand, W., 1966. On stochastic scheduling. In Yovits, M.C. et al: Research program effectiveness. Gordon and Breach, New York.

Ireland, C.T., Kullback, S., 1968. Contingency tables with given marginals. Biometrika 55, 179-188.

Kadane, J., 1966. Some equivalence classes in paired comparisons. Ann.Math. Statist. 37, 488-494.

Katz, E., 1960. Communication research and the image of society. Convergence of two traditions. Amer.J.Sociol. 65, 435.

Katz, E., Lazarsfeld, P., 1955. Personal influence. Free Press, Glencoe.

Katz, L., 1951. The distribution of the number of isolates in a social group. Ann.Math.Statist. 23, 271-276. (Reprinted in Moreno, J.L. et al. (Ed.): The sociometry reader. Free Press, Glencoe, 1960.)

Katz, L., Powell, J.H., 1954. The number of locally restricted directed graphs. Proc.Amer.Math.Soc. 5, 621-626.

Katz, L., Powell, J.H., 1957. Probability distributions of random variables associated with a structure of the sample space of sociometric investigations. Ann.Math.Statist. 28, 309-317. (Reprinted in Moreno, J.L. et al. (Ed.): The sociometry reader. Free Press, Glencoe, 1960.)

Kaufmann, A., Malgrange, Y., 1963. Recherche des chemins et circuits hamiltoniens d'un graphe. Rev. Française Recherche Opérationelle 7, 61-73.

Kelley, J.E. Jr., 1961. Critical path planning and scheduling. Mathematical basis. Operations Res. 9, 296-320.

Kemeny, J.G., Snell, J.L., 1962. Mathematical models in the social sciences. Ginn, New York.

Kendall, M.G., 1962. Rank correlation methods. Griffin, London.

Kendall, M.G., Smith, B.B., 1940. On the method of paired comparisons. Biometrika 31, 324-345.

Knuth, D.E., 1968. The art of computer programming. Vol. 1. Fundamental algorithms. Addison-Wesley, Reading.

König, D., 1936. Theorie der endlichen und unendlichen Graphen. Akad.Verl. M.B.H. Leipzig. (Reprinted by Chelsea New York, 1950.)

Krieger, M., 1967. Basic switching circuit theory. Mac Millan, New York.

Kullback, S., 1959. Information theory and statistics. Wiley, New York.

Kyburg, H.E., Smokler, H.E., (Ed.), 1964. Studies in subjective probability. Wiley, New York.

Landau, H.G., 1953. On dominance relations and the structure of animal societies. Bull.Math.Biophys. 13, 1-19.

Leavitt, H.J., 1951. Some effects of certain communication patterns on group performance. J. Abnormal Social Psychol. 46, 28-50.

Luce, R.D., 1950. Connectivity and generalized cliques in sociometric group structure. Psychometrika 15, 169-190.

Luce, R.D., Perry, A.D., 1949. A method of matrix analysis of group structure. Psychometrika 14, 95-116.

MacGrimmon, K.R., Ryavec, C.A., 1964. An analytical study of the PERT assumptions. Operations Res. 12, 16-37.

Maghout, K., 1959. Détermination des nombres de stabilité et du nombre chromatique d'un graphe. C.R.Acad.Sci. Paris 248, 3522-3523.

Maghout, K., 1963. Applications de l'algèbre de Boole à la théorie des graphes et aux programmes linéaires et quadratiques. Cahiers Centre Etud. Recherche Opérationelle 5, 21-99.

Malcolm, D.G., Roseboom, J.H., Clark, C.E., Fazar, W., 1959. Application of a technique for research and development program evaluation. Operations Res. 7, 646-670.

Moran, P.A.P., 1947. On the method of paired comparisons. Biometrika 34, 363-365.

Moreno, J.L., Jennings, H., 1938. Statistics of social configurations. Sociometry 1, 342-374. (Reprinted: Sociometric measurement of social configurations. Sociometry monograph No. 3, Beacon House, New York, 1945.)

Neyman, J., 1949. Contribution to the theory of the χ^2 test. Proc. Berkeley Symp. 239-273.

Ore, O., 1967. The four-color problem. Academic Press, New York.

Otter, R., 1948. The number of trees. Ann.Math. 49, 583-599.

Palasti, I., 1966. On the strong connectedness of directed random graphs. Studio Sci.Math. Hungar. 1, 205-214.

Parthasarathy, K.R., 1964. Enumeration of paths in digraphs. Psychometrika 29, 153-165.

Patnaik, P.B., 1949. The noncentral χ^2 and F distributions and their approximations. Biometrika 36, 202-232.

Penrose, R., 1955. A generalized inverse for matrices. Proc. Cambridge Philos.Soc. 51, 406-413.

Penrose, R., 1956. On best approximate solutions of linear matrix equations. Proc. Cambridge Philos.Soc. 52, 17-19.

Pólya, G., 1937. Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen. Acta Math. 68, 145-253.

Proctor, C.H., 1966. Two measurement and sampling methods for studying social networks. Mimeographed paper. North Carolina State University.

Proctor, C.H., 1967. Sampling social networks. Mimeographed paper for presentation at the Spring 1967 Meetings of the Southern Sociological Society.

Pyle, L.D., 1964. Generalized inverse computations using the gradient projection method. J.Assoc. Computing Machinery 11, 422-428.

Rédei, L., 1934. Ein kombinatorischer Satz. Acta Litt.Sci. (Szeged) 7, 39-43.

Remage, R., Thompson, W.A., 1966. Maximum likelihood paired comparison rankings. Biometrika 53, 143-149.

Restle, F., 1961. Psychology of judgement and choice: a theoretical essay. Wiley, New York.

Riordan, J., 1958. An introduction to combinatorial analysis. Wiley, New York.

Roach, S.A., 1968. The theory of random clumping. Methuen, London.

Roy, B., 1962. Cheminement et connexité dans les graphes. Applications aux problèmes d'ordonnancement. Metra, numéro spécial.

Sabidussi, G., 1966. The centrality index of a graph. Paper presented at the Theory of Graphs International Symposium in Rome. Publ. by Dunod, Paris, 1967.

Salomaa, A., 1969. Theory of automata. Pergamon, Oxford.

Scheffé, H., 1959. The analysis of variance. Wiley, New York.

Severo, N.C., Newman, A.E., 1960. A statistical chain-ratio method for estimating relative volumes of mail to given destinations. J.Res.Nat.Bureau Standards Section C 64, 37-47.

Shannon, C.E., 1956. The zero-error capacity of a noisy channel. I.R.E. Trans. on Information Theory 2, 8-19.

Shilov, G.E., 1961. An introduction to the theory of linear spaces. Prentice-Hall, Englewood Cliffs.

Slater, P., 1961. Inconsistencies in a schedule of paired comparisons. Biometrika 48, 303-312.

Starks, T.H., 1958. Significance tests in experiments involving paired comparisons. Virginia Polytechnical Institute.

Stephan, F.F., 1942. An iterative method of adjusting sample frequency tables when expected marginal totals are known. Ann.Math.Statist. 13, 166-178.

Stephan, F.F., 1968. Three extensions of sample survey technique. Mimeographed paper. Princeton University. Publ. in Johnson, N.L., Smith, H. Jr., (Ed.): New developments in survey sampling. Wiley, New York, 1969.

Taylor, W.F., 1953. Distance functions and regular BAN estimates. Ann.Math. Statist. 24, 85-92.

Thionet, P., 1959. L'ajustement des résultats des sondages sur ceux des dénombrements. Rev.Inst.Internat. Statist. 27, 8-25.

Thompson, W.A., Remage, R., 1964. Rankings from paired comparisons. Ann. Math.Statist. 35, 739-747.

Torgerson, W.S., 1958. Theory and methods of scaling. Wiley, New York.

Trawinski, B.J., 1961. Frequencies of partitions for paired-comparison experiments and related tables. Virginia Polytechnical Institute.

Tukey, J.W. 1957. Approximations to the upper 5 % point of Fisher's β distribution and noncentral χ^2. Biometrika 44, 528-530.

Van Slyke, R.M., 1963. Monte Carlo methods and the PERT problem. Operations Res. 11, 839-860.

Weibull, C., 1958. The distribution of reciprocal choices in sociometric tests. Statistical Institute publication 4. Göteborg University.

Zykov, A.A., 1957. Algebras of complexes. Mat.Sb.Novaja Serija 41, 159-176. (Translation by E. Buchtenkirch 1960, Amer.Math.Soc.Transl. 15, 2, 15-32.)

AUTHOR INDEX

Alway, G.G., 239
Austin, T.L., 65
Avondo-Bodino, G., 19

Barton, D.E., 66, 103
Bavelas, A., 39
Beauchamp, C., 39
Ben-Israel, A., 164
Berge, C., 19, 20, 45, 154
Berman, E.B., 19
Bhargava, T.N., 66
Blackwell, W.A., 19
Bloemena, A.R., 62, 103
Boot, J.C.G., 164
Borenius, G., 162
Bronfenbrenner, U., 220
Brown, D.T., 184
Buckner, H.T., 60

Cartwright, D., 19, 237
Chomsky, N., 20
Clark, C.E.,
(with Malcolm, D.G.), 19
Cochrun, B.L., 19
Coleman, J.S., 60
Cramér, H., 126, 248
Criswell, J.H., 220

Danzig, G.B., 19, 40
David, F.N., 66, 103
David, H.A., 67, 235, 239, 245, 252
Deming, W.E., 184
Durbin, J., 244

Edwards, D.S., 115, 220
El-Badry, M.A., 184
Erdös, P., 65

Fagen, R.E.,
(with Austin, T.L.), 65
Fazar, W.,
(with Malcolm, D.G.), 19
Feller, W., 98, 111, 207
Festinger, L., 209
Fienberg, S.E., 184
Fisk, D.L., 66
Flament, C., 19
Floyd, R.W., 19
Ford, L.R., 19, 40, 46, 154
Frank, O., 61, 66, 162, 249, 255
Fréchet, M., 207, 223

Fulkerson, D.R., 19, 46, 64, 154

Gale, D., 19, 154
Ghouila-Houri, A., 19, 154
Gilbert, E.N., 65, 211
Gill, A., 19
Ginsburg, S., 19
Goodman, L.A., 56
Graybill, F.A., 162, 163, 164, 173
Grossman, I.F., 19
Guilford, J.P., 235

Hammer, P.L., 20, 34, 46, 47, 48, 50
Harary, F., 19, 43, 53, 75, 237
Hildenbrand, W., 64

Ireland, C.T., 184

Jennings, H., 220

Kadane, J., 255
Katz, E., 60
Katz, L., 66, 115, 222, 224
Kaufmann, A., 42
Kelley, J.E. Jr., 19, 64
Kemeny, J.G., 40
Kendall, M.G., 67, 238, 245, 248
Knuth, D.E., 20
König, D., 18, 20, 196, 237
Krieger, M., 19
Kullback, S., 184, 256
Kyburg, H.E., 235

Landau, H.G., 239
Lazarsfeld, P., 60
Leavitt, H.J., 39
Lerchs, H., 19
Luce, R.D., 42, 209

MacCrimmon, K.R., 64
Maghout, K., 32, 48
Malcolm, D.G., 19
Malgrange, Y., 42
Menzel, H., 60
Moran, P.A.P., 249, 255
Moreno, J.L., 220

Newman, A.E., 156
Neyman, J., 184
Norman, R.Z., 19, 237

Ore, O., 18, 20, 49
Otter, P., 53

Palasti, I., 65
Parthasarathy, K.R., 43
Patnaik, P.B., 248
Penney, W.F., (with Austin, T.L.), 65
Penrose, R., 163
Perry, A.D., 42, 209
Pólya, G., 53, 75
Powell, J.H., 66, 222
Proctor, C.H., 60, 95
Pyle, L.D., 164

Rédei, L., 237
Remage, R., 239, 256, 257
Rényi, A., 65
Restle, F., 235
Riordan, J., 53, 75 (with Austin, T.L.), 65
Roach, S.A., 129

Roseboom, J.H.,
(with Malcolm, D.G.), 19
Rosenberg, I., 48
Ross, I.C., 43
Roy, B., 20, 33
Rudeanu, S., 20, 34, 47, 48, 50
Ryavec, C.A., 64

Sabidussi, G., 39
Salomaa, A., 20
Scheffé, H., 248
Severo, N.C., 156
Shannon, C.E., 47
Shilov, G.E., 162
Slater, P., 238, 239, 257
Smith, B.B., 67, 238, 245, 248
Smokler, H.E., 235
Snell, J.L., 40
Starks, T.H., 245
Stephan, F.F., 55, 184

Taylor, W.F., 184
Thionet, P., 184
Thompson, W.A., 239, 256, 257
Torgerson, W.S., 235
Trawinski, B.J., 245
Tukey, J.W., 248

Van Slyke, R.M., 64

Weibull, C., 66, 220, 221

Yovits, M.C.,
(with Hildenbrand, W.), 64

Zykov, A.A., 103

SUBJECT INDEX

Activity, 63
Adjacency operator, 29
Agreement, 245
Arc, 17
  circuit arc, 17, 26, 27
  directed arc, 17
  in-arc, 26
  multiple arcs, 17, 26, 27
  out-arc, 26
  single arc, 26, 27
  undirected arc, 17
Arc flow, 157
Arc-free node set, 47
Arc frequency matrix, 26, 27
Arc indicator matrix, 26, 27
Arc set, 26, 27

Bernoulli sample, 57
Bicenter, 188
Bipartite graph, 179
  complete bipartite graph, 180
Black node, 157
Boolean s-power, 32

Capacity, 46
Center, 188
  in-center, 39
  out-center, 39
Central class, 188
Chromatic decomposition, 48
Circuit arc, 17, 26, 27
Circuit path, 30
Class
  central class, 188
  connected class, 34
  strongly connected class, 35
  terminal node class, 188
  weakly connected class, 35

Clique, 209
Comparison graph, 236
Complete bipartite graph, 180
Complete component, 119
Complete graph, 119
Component
  complete component, 119
  connected component, 34
Conditional inverse, 163
Configurational analysis, 55
Connected class, 34
Connected component, 34
Connected graph, 34
Critical path, 63

Degree, 28
  in-degree, 27
  out-degree, 27
Diameter, 39
Directed arc, 17
Directed graph, 17, 26
Disjoint paths, 44
Distance, 36
  truncated distance, 37
Distance matrix, 36
Dominated node, 49
Dominating node, 49

Elementary path, 30
End node, 26

Flow
  arc flow, 157
  node flow, 157
  root flow, 197
Flow matrix, 46
Forest graph, 127

Generalized inverse, 163
Graph, 17
  bipartite graph, 179
  comparison graph, 236
  complete bipartite graph, 180
  complete graph, 119
  connected graph, 34
  directed graph, 17, 26
  forest graph, 127
  partial graph, 27
  plane graph, 49
  preference graph, 236
  simple graph, 17
  stochastic contact graph, 203, 205
  stochastic graph, 53
  subgraph, 27
  tree graph, 53, 127
  undirected graph, 17, 27
  valued graph, 17

Hamiltonian path, 237

In-arc, 26
In-center, 39
Inconsistency, 248
In-degree, 27
Independent nodes, 47
In-radius, 38
Inverse
  conditional inverse, 163
  generalized inverse, 163
Isolate, 115

Matrix
arc frequency matrix, 26, 27
arc indicator matrix, 26, 27
conditional inverse matrix, 163
distance matrix, 36
flow matrix, 46
path frequency matrix, 41
path indicator matrix, 32
projection matrix, 173
Multiple arcs, 17, 26, 27

Nearest adjoining order, 238
Nexus sampling, 55
Node, 17
black node, 157
dominated node, 49
dominating node, 49
end node, 26
independent nodes, 47
start node, 26
white node, 157
Node flow, 157
Node set, 26
arc-free node set, 47
Non-centrality parameter, 248
Null space, 162

Orthogonal projection, 163
Out-arc, 26
Out-center, 39
Out-degree, 27
Out-radius, 38

Partial graph, 27
Path, 30
circuit path, 30
critical path, 63
elementary path, 30
disjoint paths, 44
Hamiltonian path, 237
simple path, 30
Path frequency matrix, 41
Path indicator matrix, 32
Plane graph, 49
Preference graph, 236
Project, 63
Projection matrix, 173

Radius
in-radius, 38
out-radius, 38
Range space, 162
Reach, 118
Regular tree, 196
Root, 195
Root flow, 197

Sample
Bernoulli sample, 57
snowball sample, 56
s-stage sample, 57
Sampling distribution of subgraphs, 72
Score, 238
Simple graph, 17
Simple path, 30
Single arc, 26, 27
Snowball sample, 56
Social network, 60

Sociogram, 60
s-stage sample, 57
Start node, 26
Stochastic graph, 53
Stochastic contact graph, 203, 205
Stochastic transitivity, 239
Strongly connected class, 35
Structure, 29
Subgraph, 27
Subgraph total, 101

Terminal node class, 188
Total arc set, 26, 27
Tree graph, 53, 127
  regular tree, 196
Truncated distance, 37

Undirected arc, 17
Undirected graph, 17, 27

Valued graph, 17

Weakly connected class, 35
White node, 157